THE MATHEMATICS OF CIRCUIT ANALYSIS

TECHNOLOGY PRESS BOOKS

SCIENTIFIC SOCIETIES IN THE UNITED STATES
By R. S. Bates
Q. E. D., M. I. T. IN WORLD WAR II
By John Burchard
WAVELENGTH TABLES
Measured and compiled under the direction of
George R. Harrison
THE MOVEMENT OF FACTORY WORKERS
By C. A. Myers and W. R. Maclaurin
INDEX FOSSILS OF NORTH AMERICA
By H. W. Shimer and R. R. Shrock
CYBERNETICS
or Control and Communication
in the Animal and the Machine
By Norbert Wiener
EXTRAPOLATION, INTERPOLATION, AND
SMOOTHING OF STATIONARY TIME SERIES
with Engineering Applications
By Norbert Wiener

PRINCIPLES OF ELECTRICAL ENGINEERING SERIES

By Members of the Staff of the
Department of Electrical Engineering
Massachusetts Institute of Technology

ELECTRIC CIRCUITS
A First Course in Circuit Analysis for Electrical Engineers
MAGNETIC CIRCUITS AND TRANSFORMERS
A First Course for Power and Communication Engineers
APPLIED ELECTRONICS
A First Course in Electronics, Electron Tubes, and Associated Circuits
THE MATHEMATICS OF CIRCUIT ANALYSIS
Extensions to the Mathematical Training of Electrical Engineers
By E. A. Guillemin

PRINCIPLES OF
ELECTRICAL ENGINEERING
SERIES

The Mathematics of Circuit Analysis

EXTENSIONS TO THE
MATHEMATICAL TRAINING
OF ELECTRICAL ENGINEERS

by

E. A. Guillemin
PROFESSOR OF ELECTRICAL COMMUNICATIONS
DEPARTMENT OF ELECTRICAL ENGINEERING
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

A PUBLICATION OF
THE TECHNOLOGY PRESS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JOHN WILEY & SONS, INC.
NEW YORK
LONDON · CHAPMAN & HALL, LTD.

SECOND PRINTING, JANUARY 1950

PRINTED IN THE UNITED STATES OF AMERICA

To my Students
Past, present, and future

Foreword

The staff of the Department of Electrical Engineering at the Massachusetts Institute of Technology some twenty years ago undertook an extensive program of revision as a unit of its entire presentation of the basic technological principles of electrical engineering. By early 1943 this collaborative enterprise had resulted in the publication of three volumes of a projected series. Publication of this complementary book of the series was delayed by the war.

The decision to undertake so comprehensive a plan rather than to add here and patch there came from the belief that the Department's large staff, with its varied interests in teaching and related research, could effect a new synthesis of educational material in the field of electrical engineering and evolve a set of textbooks with a breadth of view not easily approached by an author working individually.

Such a comprehensive series, it was felt, should be free from the duplications, repetitions, and imbalances so often present in unintegrated series of textbooks. It should possess a unity and a breadth arising from the organization of a subject as a whole. It should be useful to the student of ordinary preparation and also provide a depth and rigor challenging to the exceptional student and acceptable to the advanced scholar. It should comprise a basic course adequate for all students of electrical engineering regardless of their ultimate specialty. Restricted to material which is of fundamental importance to all branches of electrical engineering, the course should lead naturally into any one branch.

Such a basic synthesis, it is felt, has been appropriately achieved in the first three volumes. In the course of their generation it became more and more evident that the development of further extensions of this basic material covering specialized fields would correspondingly become more and more the responsibility of individual authorities who could relate their work to the basic structure.

The four volumes and the organized program of teaching out of which they have grown, are thus the products of a major research project to improve methods of technological education. The experience gained through the years in this stimulating exploration, together with the rich background of accelerated development contributed by the circumstances of war, opens the way for further evolution in this undertaking. Perhaps the most interesting potentialities lie in the influence of the social sciences and humanities on the extensions of this vital adventure.

J. R. KILLIAN, JR.

Preface

When the writing of basic text material for the undergraduate curriculum was first undertaken by the Staff of the Department of Electrical Engineering we felt that this material should not merely fill the immediate needs of the undergraduate subjects for which it was primarily intended but in addition should offer stimulation and incentive for further study. For this reason occasional glimpses were afforded of the horizons which lie beyond the artificial boundaries set by conventional limitations in the treatment of technological principles. The collateral discussions engendered by these aims tended to appear digressive; moreover, they became so numerous as to interfere seriously with conciseness, continuity, and clarity. We therefore determined to relegate these supplemental treatments to an appendix. As the project developed the appendix became too large to justify itself; in fact, it took on the aspect of a text in itself. Thus evolved the concept of companion or reference volumes for collateral study.

Later, when we evaluated this concept as to scope and organization, we concluded that the supplemental material would be more useful if the discussions of a purely mathematical nature were collated and separated from the applications to be predicated thereon. This book is the result of this effort to avoid redundance and to attain a logical unification. The applications themselves, then, were to appear as further reference volumes of the series depending for their foundations on *The Mathematics of Circuit Analysis.*

In the course of these efforts war intervened. Through its exigencies there was an unprecedented acceleration of scientific and technological effort, compressing into a few years what normally would have taken decades to achieve. Necessarily the revision program was held in abeyance during this critical period with the result that the Department of Electrical Engineering is now reconsidering the course revision program and modifying its prewar plans. The basic series must be reappraised in the light of the vast wartime developments. This may well lead to the addition of material to the earlier basic texts. Moreover, there must be drastic reconsiderations of the supplemental volumes which will form extensions of the basic series and complement this particular reference volume. Since at the moment this book goes to press the Department is in the midst of reevaluating the extensive revision program, it is impossible here to make precise pronouncements either as to the character or

number of volumes to follow. Although thus far it has been the practice to publish the volumes of the series without naming collaborators, since the immediate volume came about largely through the inspiration of Professor Guillemin, and since it was written entirely by him under the trying pressures of war, it has been decided to recognize him as author.

As to the subject of this volume it may be well to emphasize that a mathematical textbook written by engineers is to be looked upon as an idea conveyor without claim to rigor. The discussions given herein should be regarded as being plausibility arguments rather than proofs. The primary purpose is to stimulate interest and lay a background of general understanding upon which the student may later build more carefully. In no wise therefore is the character of presentation here given to be looked upon as a substitute for the more formal rigorous treatment of the mathematician.

Although other books of this nature have been written, in the planning of this book we felt that a rather complete assemblage of mathematical topics, needed specifically or collaterally in the analysis and synthesis of electrical networks and in the attack on field problems related to transmission lines, wave guides, and antennas, was still outstanding. The opportunity to include more detailed discussions of certain topics heretofore given too little attention has been capitalized, we hope, to the advantage of student and researcher alike.

In the field of advanced algebra, for example, we believed that a discussion of determinants and matrices becomes ever so much more meaningful when coupled with the geometrical interpretations provided by the subject of linear coordinate transformations and the closely related discussion of quadratic forms. Thus the first four chapters of this book bring together a collection of topics in advanced algebra that form a closely interrelated mathematical unit, and an indispensable unit in the foundations of circuit theory or any other field of application dealing with vibrations or particle dynamics.

The fifth chapter, on vector analysis, is incorporated at this point since the geometrical and algebraic ideas involved are closely related to the foregoing material, and because the two-dimensional aspects of vector analysis and field theory are very helpful in lending physical clarity to numerous topics in complex function theory which is taken up next. Here there is considerably more detail than is to be found in existing textbooks for engineers on this subject. While the usual book for the most part is content to lay a nominal background for the methods of complex integration plus some notions about conformal transformations for field-mapping purposes, here it is our aim to provide a physical and geometrical feeling for the properties of complex functions adequate to meet diverse needs in the network synthesis field. The last two articles

in this chapter, for example, are specifically concerned with a detailed discussion of factors relevant to stability considerations and the properties of physical impedance functions.

The final chapter, dealing with Fourier series and integrals, stresses a number of items with which the engineer is especially concerned such as the convergence of Fourier series, the approximation properties of its partial sums, singularity functions and their properties, elementary transform properties, evaluation of inverse transforms through complex integration, and their approximate evaluation through use of the saddle-point method.

THE DEPARTMENT OF ELECTRICAL ENGINEERING

Cambridge, Massachusetts
January, 1949

Contents

CHAPTER IV

CHAPTER V

CHAPTER VI

PAGE

CHAPTER VII

CHAPTER I

Determinants

1. Definitions and useful properties

A discussion of the theory of determinants may be approached in a variety of ways. For the reader who already has an acquaintance with this subject and can, therefore, dispense with introductory remarks, the following procedure* is particularly effective since it strikes directly at those ideas which make the determinant a useful tool.

A determinant is commonly written in the form

$$A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \qquad [1]$$

in which the vertical lines enclosing the array of elements a_{ik} are intended to take the place of parentheses as an indication that these elements are the variables of the function A, just as $f(x)$ is written as a symbol for a function of x.

The determinant is said to be of the nth *order* when it involves n rows and n columns, the total number of elements then being n^2. The italic capital letter A is used as an abbreviation for the function whose elements are denoted by the lower case letter a. Thus, B may represent another determinant with the elements b_{ik}, etc. The first index on an element indicates the row, the second index the column in which that element is situated.

The determinant may be defined uniquely in terms of the following *three* fundamental properties:

I. *The value of the function is unchanged if the elements of any row (column) are replaced by the sums of the elements of that row (column) and the corresponding ones of another row (column); for example, if* $a_{11}, a_{12}, \cdots a_{1n}$ *are replaced by* $(a_{11} + a_{31}), (a_{12} + a_{32}), \cdots (a_{1n} + a_{3n})$.

II. *The value of the function is multiplied by the constant* k *if all the elements of any row or column are multiplied by* k.

III. *The value of the function is unity if all the elements on the principal diagonal, that is,* $a_{11}, a_{22}, \cdots a_{nn}$, *are unity and all others are zero.*

*C. Carathéodory, *Vorlesungen über reelle Funktionen* (Leipzig, 1918), Ch. VI.

To these three fundamental properties may be added the following derived ones:

IV. *The first fundamental property may be amplified to the effect that an arbitrary factor times the elements of any row or column may be added to (or subtracted from) the corresponding elements of another row or column.*

V. *The algebraic sign of the function is reversed when any two rows or columns are interchanged.*

VI. *The value of the function is zero if all the elements of a row or column are zero, or if the corresponding elements of any two rows or columns are identical or have a common ratio.*

Rule IV may be seen to follow from I and II. As shown in the numerical example below, the elements of the third column are first multiplied by k; the resulting k-multiplied elements are then added to the respective ones of the first column, after which column three is multiplied by k^{-1}, thus restoring to its elements their original values.

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 4 & 2 & 6 \\ 3 & 1 & 7 \end{vmatrix} \qquad kA = \begin{vmatrix} 1 & 3 & 2k \\ 4 & 2 & 6k \\ 3 & 1 & 7k \end{vmatrix} = \begin{vmatrix} (1+2k) & 3 & 2k \\ (4+6k) & 2 & 6k \\ (3+7k) & 1 & 7k \end{vmatrix}$$

$$A = \begin{vmatrix} (1+2k) & 3 & 2 \\ (4+6k) & 2 & 6 \\ (3+7k) & 1 & 7 \end{vmatrix} \tag{2}$$

Rule V is a consequence of I and the extended form IV of II. Thus, suppose column 1 is first added to column 3, next the resultant column 3 is subtracted from column 1, and, finally, this resulting first column is added to the resultant column 3. The net effect is to interchange columns 1 and 3 and prefix all the elements of the first column with minus signs, as illustrated below:

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 4 & 2 & 6 \\ 3 & 1 & 7 \end{vmatrix} = \begin{vmatrix} 1 & 3 & (2+1) \\ 4 & 2 & (6+4) \\ 3 & 1 & (7+3) \end{vmatrix}$$

$$= \begin{vmatrix} -2 & 3 & (2+1) \\ -6 & 2 & (6+4) \\ -7 & 1 & (7+3) \end{vmatrix} = \begin{vmatrix} -2 & 3 & 1 \\ -6 & 2 & 4 \\ -7 & 1 & 3 \end{vmatrix} \tag{3}$$

The first part of rule VI follows from the property II for $k = 0$, and the second part is seen to be true on account of IV because a row or column of zeros is obtained when, for a suitably chosen factor, the k-multiplied elements of one of the proportional rows or columns are subtracted from the respective elements of the other row or column.

2. Evaluation of numerical determinants

The properties discussed above may be applied to the numerical evaluation of determinants, as is best illustrated by the following numerical example. Let

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 4 & 2 & 6 \\ 3 & 1 & 7 \end{vmatrix} \qquad [4]$$

Step 1. Subtract from the second row the 4-multiplied elements of the first row:

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 0 & -10 & -2 \\ 3 & 1 & 7 \end{vmatrix} \qquad [5]$$

Step 2. Subtract from the third row the 3-multiplied elements of the first row:

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 0 & -10 & -2 \\ 0 & -8 & 1 \end{vmatrix} \qquad [6]$$

Step 3. Subtract from the third row the $\frac{8}{10}$-multiplied elements of the second row:

$$A = \begin{vmatrix} 1 & 3 & 2 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{vmatrix} \qquad [7]$$

Step 4. Subtract from the second column the 3-multiplied elements of the first column:

$$A = \begin{vmatrix} 1 & 0 & 2 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{vmatrix} \qquad [8]$$

Step 5. Subtract from the third column the 2-multiplied elements of the first column:

$$A = \begin{vmatrix} 1 & 0 & 0 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{vmatrix} \qquad [9]$$

Step 6. Subtract from the third column the $\frac{2}{10}$-multiplied elements of the second column:

$$A = \begin{vmatrix} 1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & \frac{13}{5} \end{vmatrix} \qquad [10]$$

Application of the fundamental properties II and III then gives

$$A = (1)(-10)(\tfrac{13}{5})\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = (1)(-10)(\tfrac{13}{5}) = -26 \qquad [11]$$

It is useful to note that the modifications involved in the last three steps of this process do not influence the values of the elements on the principal diagonal, the product of which is equal to the value of the determinant. This fact may be stated in the form of an additional derived property:

VII. *The value of the special determinant in triangular form:*

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{vmatrix} \qquad [12]$$

is given by the product $(a_{11}a_{22}\cdots a_{nn})$ *of the elements on its principal diagonal.*

With the help of this rule the value of the determinant in the above example may be set down after the completion of the third step.

If in the determinant A the rows and columns are interchanged, the values of the elements on the principal diagonal are not affected; and if the above operations with respect to rows are then replaced by the same operations with respect to corresponding columns and vice versa, the same final value is evidently arrived at. This fact demonstrates the equivalence of rows and columns as far as the value of a determinant is concerned. For convenience in reference this is stated as the property:

VIII. *The value of a determinant is unchanged if its rows are written as corresponding columns or vice versa.*

In numerical work, the method of evaluation illustrated in the above example is short and convenient to apply. When an analytic result is desired, however, other methods are usually preferable. They are given in Arts. 4 and 5, to which the discussion immediately following serves as an introduction.

3. Minors and cofactors

If in the determinant A of Eq. 1, one or more rows and a corresponding number of columns are deleted, the remaining square array of elements is again a determinant. It is referred to as the $(n - p)$-rowed *minor* (or

minor determinant) of A, where the integer p denotes the number of rows or the corresponding number of columns which have been deleted. Thus the n-rowed minor is the determinant itself. An $(n-1)$-rowed minor is also spoken of as a *first* minor, an $(n-2)$-rowed one as a *second* minor, etc.*

The minor is customarily denoted by a symbol whose indexes refer to the canceled rows and columns. Thus the minor M_{ik} is formed by canceling the ith row and the kth column in A. It is quite common to speak of M_{ik} as the minor of a_{ik}, or as the minor corresponding to the element a_{ik}, although (according to the immediately following discussion) it should more properly be referred to as the complement of a_{ik}.

A minor of second order, denoted by $M_{\substack{ik\\rs}}$, is formed by canceling the ith and rth rows and the kth and sth columns. The extension of this notation to the designation of minors of higher order is readily recognized, but when the number of canceled rows and columns is large (cases of this sort are infrequent in engineering applications), such notation becomes awkward and is usually replaced by some other expedient which seems more effective at the moment.

The elements which lie at the intersections of the canceled rows and columns, arranged in a square array in the same order (from left to right and from top to bottom) as they appear in the original determinant, form another minor determinant N which is called the *complement of* M. The complement of a first minor is a single element; that of a second minor is a two-rowed determinant, etc.

In particular, the minors formed by canceling the *same* rows as columns (these intersect on the principal diagonal) are called *principal* minors, and their complements are again principal minors.

An alternative view may be taken with regard to the formation of minors. Instead of obtaining the minors by canceling rows and columns in the original determinant, they may be formed by first selecting from the determinant certain rows, and subsequently selecting from this rectangular (nonsquare) array any like number of columns. Or, from a given set of columns in the determinant A, minors may be formed by the selection of corresponding numbers of rows. The minors thus formed are evidently the complements of those obtained by the process of canceling the same combinations of rows and columns.

If M is a minor of the determinant A and N is its complement, and if the rows and columns *contained* in M are formed from the $i, j, k, \cdots$ th rows and the $q, r, s, \cdots$ th columns of A, then

$$N \times (-1)^{(i+j+k+\cdots+q+r+s+\cdots)} \qquad [13]$$

*The terms "minor of first order," "second order," etc., are also used.

is referred to as the *algebraic* complement of M. This differs from the ordinary complement only in its algebraic sign. If the sum of the indexes referring to the (first, second, etc.) rows and columns of A contained in M is an even integer, the algebraic sign, by Eq. 13, is $+1$; if this sum is an odd integer, the sign is -1.

The relation between a minor and its complement is evidently a mutual one in the sense that the two designations may be interchanged. Whereas M may be called a minor and N its complement, N may be looked upon as the minor and M as its complement. Thus, a single element may be thought of as a one-rowed minor. If the element is a_{ik}, its complement is the minor M_{ik}.

The algebraic complement of the single element a_{ik} is sufficiently important to deserve a special name and symbol. It is called the *cofactor* of a_{ik} and is quite commonly denoted by the corresponding upper-case letter with like subscripts (although various other notations are also encountered in the literature). Thus, the cofactor of a_{ik} is given by

$$A_{ik} = (-1)^{i+k} M_{ik} \tag{14}$$

It differs from the minor (which is the complement of a_{ik}) only in algebraic sign; hence the cofactor is sometimes referred to as the *signed* minor.

The indexes i and k, whose integer values determine the sign-controlling factor $(-1)^{i+k}$, refer respectively to the row and column intersecting at the point where the element a_{ik} is located. If the cofactor is formed for an element in the original determinant, its indexes and those appearing in the sign-controlling factor obviously agree with the indexes appearing on the element in question, because the indexes on an element of the original determinant indicate respectively its row and its column positions. This correspondence is, however, no longer consistently true in a minor of the original determinant.

For example, the minor M_{23} of A, Eq. 1, reads

$$M_{23} = \begin{vmatrix} a_{11} & a_{12} & a_{14} & a_{15} & \cdots & a_{1n} \\ a_{31} & a_{32} & a_{34} & a_{35} & \cdots & a_{3n} \\ a_{41} & a_{42} & a_{44} & a_{45} & \cdots & a_{4n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n4} & a_{n5} & \cdots & a_{nn} \end{vmatrix} \tag{15}$$

Here the element a_{32}, for instance, is located at the intersection of the second row and the second column. This is referred to as the (2,2) position in the minor determinant of Eq. 15. In general, the term (r,s) *position* is used to indicate the location at which the rth row and sth column of a given rectangular array intersect. The object of the present argument is to point out as a typical case that in forming the cofactor for the element

a_{32} for the minor determinant of Eq. 15, the sign-controlling factor is $(-1)^{2+2}$ and not $(-1)^{3+2}$.

If only the algebraic signs of the cofactors are set down at the positions of the corresponding elements in a rectangular array, the following checkerboard of $+$ and $-$ signs is obtained:

$$\begin{vmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{vmatrix} \qquad [16]$$

This pictorial statement for the signs of the cofactors is sometimes referred to as the "checkerboard rule."

4. Laplace's development

In the manipulation of determinants, and sometimes to facilitate their numerical evaluation, a process of development formulated by Laplace is frequently useful. It may be stated in the following form:

If all the minors are formed from a selected set of rows or columns of a determinant and the products of these minors with their respective algebraic complements are added, the resulting sum is equal to the determinant.

If a single row is selected, this development reads

$$A = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in} \qquad (i = 1, \ 2, \ \cdots n) \quad [17]$$

For a single column, the result is written

$$A = a_{1k}A_{1k} + a_{2k}A_{2k} + \cdots + a_{nk}A_{nk} \qquad (k = 1, \ 2, \ \cdots n) \quad [18]$$

In Eq. 17 the determinant is represented by the sum of the products of the elements of any row with their respective cofactors. In Eq. 18 a corresponding summation is carried out with respect to the elements and cofactors of any column. This simplest form for the Laplace development, which is also called an expansion of the determinant along one of its rows or columns, is the one most frequently used.

It may be of interest, however, to illustrate a more complicated example of this type of development. Let the following fourth-order determinant

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} \qquad [19]$$

be developed through the selection of the first two columns for the forma-

tion of minors. All possible two-rowed minors are systematically formed as the rows: 1, 2; 1, 3; 1, 4; 2, 3; 2, 4; 3, 4 are selected from these columns. The sign-controlling factors of the corresponding algebraic complements are respectively:

$$
\begin{aligned}
(-1)^{1+2+1+2} &= +1\\
(-1)^{1+2+1+3} &= -1\\
(-1)^{1+2+1+4} &= +1\\
(-1)^{1+2+2+3} &= +1\\
(-1)^{1+2+2+4} &= -1\\
(-1)^{1+2+3+4} &= +1
\end{aligned}
$$

With the terms written down in this order, the development reads

$$
\begin{aligned}
A = {}& \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \times \begin{vmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{vmatrix} - \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \times \begin{vmatrix} a_{23} & a_{24} \\ a_{43} & a_{44} \end{vmatrix} \\
&+ \begin{vmatrix} a_{11} & a_{12} \\ a_{41} & a_{42} \end{vmatrix} \times \begin{vmatrix} a_{23} & a_{24} \\ a_{33} & a_{34} \end{vmatrix} + \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \times \begin{vmatrix} a_{13} & a_{14} \\ a_{43} & a_{44} \end{vmatrix} \\
&- \begin{vmatrix} a_{21} & a_{22} \\ a_{41} & a_{42} \end{vmatrix} \times \begin{vmatrix} a_{13} & a_{14} \\ a_{33} & a_{34} \end{vmatrix} + \begin{vmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{vmatrix} \times \begin{vmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{vmatrix}
\end{aligned} \qquad [20]
$$

By means of the Laplace development a determinant may evidently be evaluated in a variety of ways. One possible method of evaluation consists in repeatedly applying the simplest form of development given by Eqs. 17 and 18. In the first step of this process, the development is given by the sum of n terms, each of which is the product of an element and an $(n-1)$-rowed cofactor. In the second step, each of these cofactors is similarly developed, thus yielding for the determinant A a sum of $n(n-1)$ terms, each of which consists of the product of two elements and an $(n-2)$-rowed cofactor. As the process is continued one recognizes that the final evaluation of A is given by the sum of $n!$ terms, each of which consists of the product of n elements.

The determinant is, therefore, a rational integral function, homogeneous, and of the nth degree in its elements. In any term of the final evaluated form, the appearance of the product of an element with another element of the *same* row or column is not possible. This fact is readily appreciated by noting in the term $a_{12}A_{12}$, for example, that the cofactor A_{12} contains no elements of the first row or second column. Hence none of these elements can subsequently appear in a term containing a_{12}. The determinant is, therefore, a *linear* function in the elements of any one row or column.

5. Other methods of evaluation in numerical or functional form

The evaluation of a determinant by means of the Laplace development, although useful for numerous analytic investigations, is a long and tedious process. The solution of simultaneous linear equations by means of determinants, as discussed in Art. 8, is usually found in numerical problems to involve a larger number of component operations than a systematic process of elimination. This situation is true even when the determinant and cofactors are evaluated by the method given in Art. 2, although this method parallels the elimination process in its essential steps.

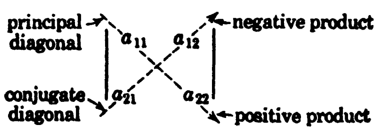
Alternative abbreviated methods of solving such equations are given in Arts. 7 and 11 of Ch. II. From a broader standpoint it is well to be familiar with numerous processes of evaluating determinants, so that the particular conditions of a specific problem may be met most expeditiously. In this regard the following remarks may also prove useful.

Evaluations of two of the simplest cases by means of the Laplace development method are written down so that their resultant forms may be examined.

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12} \tag{21}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{matrix} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{33}a_{21}a_{12} \end{matrix} \tag{22}$$

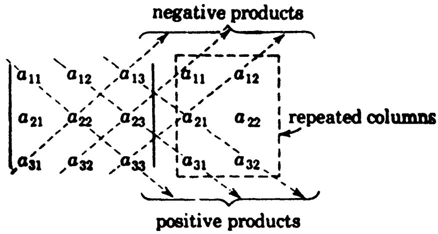
By inspection of Eq. 21 it may be said that the value of a two-rowed determinant is given by the product of the elements on the principal diagonal less the product of the elements on the conjugate diagonal (lower left to upper right) as indicated in the following by arrows:



[23]

This is called the ***diagonal product rule***. It is applicable in extended form to the evaluation of a three-rowed determinant. Here there are three positive and three negative products, the positive ones being formed by elements on the principal and adjacent parallel diagonals and the negative ones by elements on the conjugate and adjacent parallel diagonals in a

manner which is more easily understood if the first two columns are repeated so that the arrows may continue straight, thus:



[24]

The result is seen to check with Eq. 22.

An extension of this rule does not yield the value of fourth and higher order determinants, as may readily be appreciated from the fact that the number of terms in the final evaluation must be $n!$, whereas the diagonal product rule yields only $2n$ terms. If $n = 4$, there remain $4! - 2 \times 4 = 16$ terms unaccounted for after all diagonal products have been formed.

From a more comprehensive study of determinants, it is seen that all the terms in the final evaluation may be found by writing down the group of elements on the principal diagonal

$$a_{11} \quad a_{22} \quad a_{33} \cdots a_{nn}$$

and carrying out all permutations of the first subscripts, keeping the second subscripts fixed, or vice versa. In either case there are as many different products as there are permutations of n things taken n at a time, which is $n!$

In this process, the algebraic signs of the various terms are controlled by the rule that all permutations formed by an *even* number of inversions of the permuted subscripts represent *positive* terms, all others being negative. Thus for $n = 4$ the possible permutations are

Even number of inversions	*Odd number of inversions*
1 2 3 4	2 1 3 4
2 3 1 4	2 3 4 1
1 3 4 2	1 3 2 4
3 1 2 4	3 1 4 2
3 4 1 2	3 2 1 4
3 2 4 1	3 4 2 1
2 1 4 3	1 2 4 3
2 4 3 1	2 4 1 3
1 4 2 3	1 4 3 2
4 1 3 2	4 1 2 3
4 2 1 3	4 3 1 2
4 3 2 1	4 2 3 1

The twenty-four terms written with these as the first or second set of indexes, the fixed set being 1 2 3 4, and prefixed with algebraic signs according to the stated rule, represent the evaluation of the fourth-order determinant.

6. Bordered determinants

A determinant of given order may be transformed into a determinant of higher order without changing its value, as is readily seen by applying the ideas of the Laplace development to the following example:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 1 & \alpha_0 & \alpha_1 & \alpha_2 \\ 0 & 1 & \beta_1 & \beta_2 \\ 0 & 0 & a_{11} & a_{12} \\ 0 & 0 & a_{21} & a_{22} \end{vmatrix} \qquad [25]$$

The elements α_0, α_1, α_2, β_1, β_2 may have any values. The process can evidently be varied by placing the zeros to the right or above or below the rectangle containing the a_{ik}'s. The resulting form is referred to as a *bordered* determinant.

7. Products of determinants

The product of two determinants of like order can be expressed as a single determinant of the same order. If the two determinants are initially not of the same order, one of them can be bordered. In the present discussion the determinants can, therefore, be assumed to have the same order.

The procedure for obtaining the elements of the product determinant is best illustrated by means of a simple example. By the Laplace development the transformation of the following product is justified:

$$A \times B = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \times \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & -1 & 0 \\ a_{21} & a_{22} & 0 & -1 \\ 0 & 0 & b_{11} & b_{12} \\ 0 & 0 & b_{21} & b_{22} \end{vmatrix} \qquad [26]$$

According to rule IV, Art. 1, the resultant fourth-order determinant may be modified in the following manner without changing its value. First the b_{11}-multiplied elements of the first row and the b_{12}-multiplied elements of the second row are added to the corresponding elements of the third row, giving

$$A \times B = \begin{vmatrix} a_{11} & a_{12} & -1 & 0 \\ a_{21} & a_{22} & 0 & -1 \\ c_{11} & c_{21} & 0 & 0 \\ 0 & 0 & b_{21} & b_{22} \end{vmatrix} \qquad [27]$$

in which

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{21}b_{12} \\ c_{21} &= a_{12}b_{11} + a_{22}b_{12} \end{aligned} \qquad [28]$$

The object of this transformation is to produce zeros in place of the elements b_{11} and b_{12} in the fourth-order determinant of Eq. 26. Now both the b_{21}-multiplied elements of the first row and the b_{22}-multiplied elements of the second row are added to the corresponding elements of the fourth row, giving

$$A \times B = \begin{vmatrix} a_{11} & a_{12} & -1 & 0 \\ a_{21} & a_{22} & 0 & -1 \\ c_{11} & c_{21} & 0 & 0 \\ c_{12} & c_{22} & 0 & 0 \end{vmatrix} \qquad [29]$$

where

$$\begin{aligned} c_{12} &= a_{11}b_{21} + a_{21}b_{22} \\ c_{22} &= a_{12}b_{21} + a_{22}b_{22} \end{aligned} \qquad [30]$$

By the method of Laplace's development the determinant of Eq. 29 is simply

$$A \times B = \begin{vmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \end{vmatrix} = \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix} \qquad [31]$$

the second form being obtained by an interchange of rows and columns. Examining the expressions for the elements c_{ik}, as given by Eqs. 28 and 30, it is observed that they are formed by multiplying successive elements in the columns of the determinant A by successive elements in the rows of the determinant B and adding the results, the specific columns and rows involved being indicated by the first and second subscripts respectively on c_{ik}. Thus c_{11} is formed from the elements of the first column of A and those of the first row of B; c_{12} is formed from the elements of the first column of A and those of the second row of B, etc. More briefly, the c's are said to be formed by multiplying the columns of A by the rows of B.

Since the individual values of the determinants A and B are unchanged by writing their rows as columns, it is clear that the value of the product determinant is unaltered if its elements are formed by multiplying either the rows *or* columns of A by either the rows *or* columns of B. The elements of this resulting determinant may, therefore, be formed in any of *four* different ways. Although the individual elements thus obtained are different, the value of the resultant determinant remains the same.

A straightforward extension of the method used in the above example shows that the process of forming the elements of a determinant representing the product of two given determinants A and B follows the same general rules when A and B have any order. This result is summarized in the statement

$$\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} \times \begin{vmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{vmatrix} = \begin{vmatrix} c_{11} & \cdots & c_{1n} \\ \cdots & \cdots & \cdots \\ c_{n1} & \cdots & c_{nn} \end{vmatrix} \qquad [32]$$

with any one of the following four relations:

$$c_{ik} = a_{i1}b_{k1} + a_{i2}b_{k2} + \cdots + a_{in}b_{kn} \qquad [33]$$

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk} \qquad [34]$$

$$c_{ik} = a_{1i}b_{k1} + a_{2i}b_{k2} + \cdots + a_{ni}b_{kn} \qquad [35]$$

$$c_{ik} = a_{1i}b_{1k} + a_{2i}b_{2k} + \cdots + a_{ni}b_{nk} \qquad [36]$$

On the right-hand sides of the last four expressions, the indexes i and k have fixed values for any one of the c's. Since the first index refers to a row and the second to a column, it is recognized that the four representations for c_{ik} above correspond respectively to multiplying rows of A by rows of B, rows of A by columns of B, columns of A by rows of B, and columns of A by columns of B.

8. Linear Equations

One of the most important uses for determinants is in the solution of linear simultaneous equations. A set of n such equations involving n unknowns reads

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{array}\right\} \qquad [37]$$

The object is to determine the values $x_1 \cdots x_n$ in terms of the values $y_1 \cdots y_n$ and the coefficients a_{ik} which are the elements of the determinant of this system of equations. This determinant is given by Eq. 1.

Let the equations in the set (Eq. 37) be multiplied successively by the cofactors $A_{11}, A_{21}, \cdots + A_{n1}$. Subsequent addition yields the equation

$$\begin{array}{l} (a_{11}A_{11}+a_{21}A_{21}+\cdots+a_{n1}A_{n1})x_1 \\ +(a_{12}A_{11}+a_{22}A_{21}+\cdots+a_{n2}A_{n1})x_2 \\ +(a_{13}A_{11}+a_{23}A_{21}+\cdots+a_{n3}A_{n1})x_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ +(a_{1n}A_{11}+a_{2n}A_{21}+\cdots+a_{nn}A_{n1})x_n = A_{11}y_1+A_{21}y_2+\cdots+A_{n1}y_n \end{array} \qquad [38]$$

Here the coefficient of x_1 is recognized as the Laplace development of the determinant A with respect to the elements of the first column, as given by Eq. 18 for $k = 1$. Similarly, the coefficient of x_2 is seen to be the Laplace development, with respect to the elements of the first column, of a determinant which is formed from A by replacing the elements of the

first column by those of the second column. This coefficient, therefore, is the determinant

$$\begin{vmatrix} a_{12} & a_{12} & a_{13} & \vdots & a_{1n} \\ a_{22} & a_{22} & a_{23} & \vdots & a_{2n} \\ \cdot & \cdot & \cdot & \vdots & \cdot \\ \cdot & \cdot & \cdot & \vdots & \cdot \\ \cdot & \cdot & \cdot & \vdots & \cdot \\ a_{n2} & a_{n2} & a_{n3} & \vdots & a_{nn} \end{vmatrix} \tag{39}$$

which by rule VI, Art. 1, has the value zero.

Similarly, the coefficient of x_3 is equal to the determinant A with its first column replaced by the third column. This is likewise zero, as are the coefficients of all the remaining x's. Equation 38 is, therefore, equivalent to

$$Ax_1 = A_{11}y_1 + A_{21}y_2 + \cdots + A_{n1}y_n \tag{40}$$

whence

$$x_1 = \frac{A_{11}y_1 + A_{21}y_2 + \cdots + A_{n1}y_n}{A} \tag{41}$$

In like manner the solution for x_2 may be obtained by multiplying the equations in the set 37 by the cofactors $A_{12}, A_{22}, \cdots A_{n2}$, respectively and adding the results. The coefficient of x_2 then equals A, and the remaining ones are zero. Hence there results

$$x_2 = \frac{A_{12}y_1 + A_{22}y_2 + \cdots + A_{n2}y_n}{A} \tag{42}$$

This result may be stated in general terms by assuming Eqs. 37 to be multiplied respectively by the cofactors $A_{1k}, A_{2k}, \cdots + A_{nk}$ and adding the results. The coefficient of x_k then equals A, and the remaining ones are zero, so that

$$x_k = \frac{A_{1k}y_1 + A_{2k}y_2 + \cdots + A_{nk}y_n}{A} \tag{43}$$

For $k = 1, 2, \cdots n$, this represents the desired solutions.

The numerator of Eq. 43 is recognized as the Laplace development of a determinant which is formed from A by replacing its kth column by the column of y's appearing on the right-hand sides of Eqs. 37. Thus the result, Eq. 43, may be written

$$x_k = \frac{\begin{vmatrix} a_{11} & \vdots & a_{1,k-1} & y_1 & a_{1,k+1} & \vdots & a_{1n} \\ a_{21} & \vdots & a_{2,k-1} & y_2 & a_{2,k+1} & \vdots & a_{2n} \\ \cdot & \vdots & \cdot & \cdot & \cdot & \vdots & \cdot \\ \cdot & \vdots & \cdot & \cdot & \cdot & \vdots & \cdot \\ \cdot & \vdots & \cdot & \cdot & \cdot & \vdots & \cdot \\ a_{n1} & \vdots & a_{n,k-1} & y_n & a_{n,k+1} & \vdots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}} \tag{44}$$

A statement describing this form of the solution is known as *Cramer's rule*.

A significant feature in the derivation of these solutions is a recognition of the validity of the relation

$$a_{1i}A_{1k} + a_{2i}A_{2k} + \cdots + a_{ni}A_{nk} = \begin{cases} A & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \tag{45}$$

which justifies the step from Eq. 38 to Eq. 40. The companion relation, which is established in an analogous fashion, reads

$$a_{i1}A_{k1} + a_{i2}A_{k2} + \cdots + a_{in}A_{kn} = \begin{cases} A & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \tag{46}$$

For $i \neq k$ this represents the Laplace development of a determinant whose ith and kth *rows* are identical. Equations 45 and 46 may he looked upon as an extension of the relations expressed by Eqs. 18 and 17 respectively.

The solutions to the set of Eqs. 37 may be written in the form

$$\left.\begin{aligned} b_{11}y_1 + b_{12}y_2 + \cdots + b_{1n}y_n &= x_1 \\ b_{21}y_1 + b_{22}y_2 + \cdots + b_{2n}y_n &= x_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\ b_{n1}y_1 + b_{n2}y_2 + \cdots + b_{nn}y_n &= x_n \end{aligned}\right\} \tag{47}$$

in which, according to Eq. 43, the coefficients are given by

$$b_{rs} = \frac{A_{sr}}{A} \tag{48}$$

In this result it is significant to note the reversal of the subscripts on A_{sr} as compared with those on b_{rs}.

In case the elements of the determinant A fulfill the condition

$$a_{ik} = a_{ki} \tag{49}$$

the determinant is said to be *symmetrical* about its principal diagonal. It is clear from rule VIII, Art. 2, that the minors and cofactors of A then

also have this property, that is,

$$A_{ik} = A_{ki} \tag{50}$$

and it then follows from Eq. 48 that the determinant of the system of Eqs. 47 is likewise symmetrical. In that case the subscript order in Eq. 48 is, of course, unimportant.

Equations 37 and 47 are mutually inverse systems. The one set represents the solution of the other. Consequently by analogy to Eq. 48 the coefficients of Eqs. 37 may be written

$$a_{ik} = \frac{B_{ki}}{B} \tag{51}$$

in which

$$B = \begin{vmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{vmatrix} \tag{52}$$

is the determinant of the system of Eqs. 47 and B_{ki} its cofactors.

Evaluating the product AB of the determinants of the inverse systems of Eqs. 37 and 47 by substituting Eq. 48 into Eq. 34 gives

$$c_{ik} = \frac{a_{i1}A_{k1} + a_{i2}A_{k2} + \cdots + a_{in}A_{kn}}{A} \tag{53}$$

Reference to Eq. 46 shows that

$$c_{ik} = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \tag{54}$$

Hence for the determinants of these inverse systems

$$\begin{vmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{vmatrix} \times \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{vmatrix} = 1 \tag{55}$$

The determinants have inverse values.

9. Conditions for the existence of solutions

The conditions under which a system of simultaneous equations such as the set 37 can have solutions may be seen from the form of these solutions as expressed by Eq. 44. For arbitrary values of $y_1 \cdots y_n$, a necessary and sufficient condition for the existence of these solutions is that the determinant A shall not be zero. If A is zero, in general no solutions exist.

They may, however, still exist in case the determinant in the numerator of Eq. 44 is also zero, as it is if the y's satisfy the conditions

$$y_i = \alpha_1 a_{i1} + \alpha_2 a_{i2} + \cdots + \alpha_n a_{in} \tag{56}$$

in which $\alpha_1, \cdots \alpha_n$ are arbitrary factors. The column of y's in the numerator of Eq. 44 is then a linear combination of the other columns of this determinant and can, by repeated application of rule IV, Art. 1, be reduced to a column of zeros.

When the y's are expressed by the relations 56, the Eqs. 37 can be rewritten in the form

$$\left.\begin{array}{l} a_{11}(x_1 - \alpha_1) + a_{12}(x_2 - \alpha_2) + \cdots + a_{1n}(x_n - \alpha_n) = 0 \\ a_{21}(x_1 - \alpha_1) + a_{22}(x_2 - \alpha_2) + \cdots + a_{2n}(x_n - \alpha_n) = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}(x_1 - \alpha_1) + a_{n2}(x_2 - \alpha_2) + \cdots + a_{nn}(x_n - \alpha_n) = 0 \end{array}\right\} \qquad [57]$$

A special case of this sort occurs when all the y's are zero. Then $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$, and Eqs. 57 become identical with Eqs. 37 for $y_1 = y_2 = \cdots = y_n = 0$. This is called the corresponding *homogeneous* set of equations. For these, the Cramer rule as expressed by Eq. 44 yields the solutions in indeterminate form except when $A \neq 0$, but then the solutions are all zero. They are spoken of as the trivial solutions because their existence is at once evident upon inspection of the homogeneous equations.

Nontrivial solutions to the homogeneous set of Eqs. 57 exist only if the determinant is zero, but Cramer's rule, Eq. 44, is of no use in determining them. In order to see how this difficulty might be overcome, it is helpful to consider the Eqs. 37 for the special case that one of the y's, for example, y_i, alone is different from zero. Then, if it is assumed for the moment that the determinant is not zero and Cramer's rule or Eq. 43 is applied, it is found that

$$x_k = \frac{A_{ik}y_i}{A} \qquad [58]$$

In this special case the ratio of any two unknowns is given by

$$\frac{x_r}{x_s} = \frac{A_{ir}}{A_{is}} \qquad [59]$$

which is independent of the values of both y_i and the determinant A. It may be inferred, therefore, that Eq. 59 holds also when both y_i and A are zero.

The correctness of this conclusion is demonstrated in a rigorous fashion in Art. 7, Ch. III. Meanwhile it is interesting to note that when the homogeneous set of equations has nontrivial solutions, these are not uniquely determined by Eq. 59, in which only the ratios of the unknowns are given. Any value can be assigned to one of them, and the remaining unknowns are then expressed in terms of this one.

The general conditions for the existence of solutions may be discussed as follows. The fact that the inhomogeneous equations 37 have solutions only when the determinant is not zero simply amounts to stating that these equations must be independent. If one is a linear combination of the others (in this case the determinant vanishes), then, speaking in physical terms, the data are insufficient to yield an explicit answer.

In case the right-hand members of the Eqs. 37 are zero and all the equations are independent ($A \neq 0$), the system is overspecified from a physical point of view. The situation is like a deadlock, and nothing can happen; that is, only zero values for the unknowns can satisfy the equations. If one of the equations is a linear combination of the others ($A = 0$), this one may be discarded and one of the terms in each remaining equation, for example, that with x_n, transposed to the right-hand side. These $(n - 1)$ equations may then be solved for the $(n - 1)$ remaining unknowns in terms of x_n, provided the determinant of this reduced set is not zero. If it is zero, this method fails, but so does the corresponding form of solution expressed by Eq. 59.

This kind of failure in the method of solution indicates that two independent sets of solutions exist, but it is difficult to obtain a clear picture of this situation without the aid of such appropriate geometrical interpretations as are given in Ch. III. The present discussion is completed in that chapter. The material of the following article, however, is helpful in summarizing some of the characteristics of the determinant which are pertinent to the present problem.

10. The rank of a determinant

If in the determinant A, Eq. 1, there exists among the elements of each column the *same* linear relation

$$\alpha_1 a_{1k} + \alpha_2 a_{2k} + \cdots + \alpha_n a_{nk} = 0 \qquad (k = 1, 2, \cdots n) \tag{60}$$

in which the α's are arbitrary nonzero factors, the elements of any row are expressible as linear combinations of the corresponding elements of the remaining rows. If some of the factors $\alpha_1 \cdots \alpha_n$ are zero, this fact still holds for the elements of some of the rows. By repeated modification of the determinant according to rule IV, Art. 1, any one of these rows can be reduced to a row of zeros. Hence it is seen that a determinant is zero if a relation of the form given by Eq. 60 exists in which at least one of the α's is different from zero.

Conversely, if the determinant is known to be zero, it is surely possible to find a relation of the form of Eq. 60, as is clear if Eq. 60 is written out for all the k-values, thus

$$\left.\begin{array}{l} a_{11}\alpha_1 + a_{21}\alpha_2 + \cdots + a_{n1}\alpha_n = 0 \\ a_{12}\alpha_1 + a_{22}\alpha_2 + \cdots + a_{n2}\alpha_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{1n}\alpha_1 + a_{2n}\alpha_2 + \cdots + a_{nn}\alpha_n = 0 \end{array}\right\} \qquad [61]$$

This is simply the homogeneous system of equations in which the determinant has its columns written as rows. It is called the *transposed* set of equations. According to property VIII, Art. 2, the transposed determinant is zero if the given one is. Hence, by the method discussed in the previous article, a set of nonzero α-values can be found to satisfy Eqs. 61.

It is then also possible to find a set of nonzero β-values satisfying the relations

$$\beta_1 a_{i1} + \beta_2 a_{i2} + \cdots + \beta_n a_{in} = 0 \qquad (i = 1, 2, \cdots n) \qquad [62]$$

since they are simply a set of solutions to the untransposed system of homogeneous equations.

It follows that if the elements in a row of a determinant are expressible as linear combinations of the corresponding elements of the remaining rows, the elements in a column of this determinant are similarly expressible.

When a determinant of order n is not zero, so that neither the relations 60 nor 62 exist (except for zero α- and β-values), it is said to have the *rank* $r = n$. If the determinant is zero but at least one of its $(n - 1)$-rowed minors is different from zero, one set of α- and β-values exist which satisfy the Eqs. 60 and 62. The determinant is then said to have the rank $r = n - 1$. If the determinant and *all* its $(n - 1)$-rowed minors are zero but at least one of its $(n - 2)$-rowed minors is different from zero, it has the rank $r = n - 2$. In this case it is possible to find *two* independent sets of α- and β-values satisfying Eqs. 60 and 62, and the homogeneous system of equations has two independent sets of solutions, as is discussed further in Ch. III. Corresponding statements for the condition that the rank is $n - 3$, $n - 4$, etc., are obvious. Finally, if the rank of a determinant is zero, all its elements are zero.

The definition of rank facilitates the discussion of possible solutions to a system of linear equations. Thus the nonhomogeneous Eqs. 37 have solutions only when the rank of their determinant is n. The corresponding homogeneous equations have nontrivial solutions only if the rank is less than n, and they have p sets of independent solutions if the rank is $n - p$. A more lucid exposition of this line of reasoning is found in the chapters immediately following.

PROBLEMS

1. Determine the rank of each of the following determinants:

$$\text{(a)}\ \begin{vmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{vmatrix} \quad \text{(b)}\ \begin{vmatrix} 3 & 9 & 20 & 18 \\ 8 & 19 & 40 & 37 \\ 13 & 20 & 47 & 34 \\ 20 & 22 & 59 & 31 \end{vmatrix} \quad \text{(c)}\ \begin{vmatrix} 5 & 0 & -2 & 9 \\ 0 & 5 & -11 & 7 \\ -2 & -11 & 25 & -19 \\ 9 & 7 & -19 & 26 \end{vmatrix}$$

$$\text{(d)}\ \begin{vmatrix} 12 & -5 & 8 \\ 4 & 0 & 1 \\ -4 & 3 & -4 \end{vmatrix} \quad \text{(e)}\ \begin{vmatrix} 39 & 24 & 12 & 5 \\ 24 & 21 & 2 & 2 \\ 12 & 2 & 10 & 3 \\ 5 & 2 & 3 & 1 \end{vmatrix} \quad \text{(f)}\ \begin{vmatrix} 6.5 & 6.5 & 4 \\ 6.5 & 17 & 8 \\ 4 & 8 & 4 \end{vmatrix}$$

2. Transform each of the above determinants to the triangular form, thus finding their values and checking the answers to Prob. 1.

3. For each determinant in Prob. 1 whose rank is less than its order, find relations of the form of Eqs. 60 and 62.

4. Using determinants (e) and (f) of Prob. 1, write down corresponding sets of simultaneous equations, denoting the right-hand members by $y_1, y_2, \cdots$ as in Eqs. 37. Solve these equations by means of Cramer's rule.

5. Repeat the solutions of the equations in Prob. 4 by means of a systematic elimination process. Compare the total number of multiplications and additions with those required in the solutions using Cramer's rule.

6. Evaluate the following determinant according to the pattern shown in Eq. 20.

$$\begin{vmatrix} 2 & 1 & 4 & 3 \\ 6 & -1 & 2 & -4 \\ 3 & -2 & 5 & 1 \\ -5 & 6 & 4 & -1 \end{vmatrix}$$

Repeat the evaluation through reduction to the triangular form and compare the total numbers of multiplications and additions required in the two methods. Derive a formula giving the total numbers of multiplications and of additions required for the evaluation of an nth order determinant by the method involving its reduction to the triangular form.

7. Determine the solutions to a set of equations (like Eqs. 37) having the ac-

$$\begin{vmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ -0.866 & 0.289 & 0.289 & 0.289 \\ 0 & -0.258 & 0.408 & 0.408 \\ 0 & 0 & -0.707 & 0.707 \end{vmatrix}$$

companying determinant. Compare the set of equations representing the solutions with the given equations and note any obvious mutual relations existing between these two sets of equations.

8. Given the two sets of equations

$$\sum_{k=1}^{n} a_{ik}x_k = y_i \qquad (i = 1, 2, \cdots n)$$

and

$$\sum_{s=1}^{n} b_{ks}y_s = x_k \qquad (k = 1, 2, \cdots n)$$

which are solutions of each other, show that the corresponding determinants A and B have inverse values; that is, $AB = 1$. A proof may be based upon the rule for forming the product of two determinants.

9. In the following special nth order determinant

$$D_n = \begin{vmatrix} \alpha & 1 & 0 & 0 \cdots & 0 \\ 1 & \alpha & 1 & 0 \cdots & 0 \\ 0 & 1 & \alpha & 1 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 \cdots & 0 & 1 & \alpha \end{vmatrix}$$

all the elements of the principal diagonal are equal to α; those on the diagonals immediately above and below the principal diagonal are unity, and all the rest are zero. Derive the following recursion formula:

$$D_n = \alpha D_{n-1} - D_{n-2}$$

applicable for $n = 1, 2, \cdots$ with the definitions: $D_0 = 1$ and $D_1 = 0$. From this recursion formula obtain an explicit expression for the determinant of order n which reads

$$D_n = \frac{\sinh\,(n+1)\gamma}{\sinh\gamma}, \quad \text{with } \gamma = \cosh^{-1}\frac{\alpha}{2}$$

10. If the first and the last elements on the principal diagonal of the determinant in Prob. 9 are replaced by $\alpha\;2$, show that the resulting determinant has the value

$$D_n = \sinh\,(n-1)\gamma \cdot \sinh\gamma$$

while if *only the first or last* of these elements is $\alpha/2$, the value is given by

$$D_n = \cosh n\gamma$$

11. Consider the determinant

$$D = \begin{vmatrix} d_1 & 1 & 0 & 0 & 0 & 0 \cdots \\ -1 & d_2 & 1 & 0 & 0 & 0 \cdots \\ 0 & -1 & d_3 & 1 & 0 & 0 \cdots \\ 0 & 0 & -1 & d_4 & 1 & 0 \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{vmatrix} = K(d_1, \cdots d_n)$$

and show that this function $K(d_1, \cdots d_n)$, called a simple *continuant*, possesses the recursion formula

$$K(d_1, \cdots d_n) = d_n K(d_1, \cdots d_{n-1}) + K(d_1, \cdots d_{n-2})$$

with

$$K(d_1) = d_1, \qquad \text{and} \qquad K(0) = 1$$

12. Denote by D_{11} the cofactor formed through canceling the first row and column in the determinant given in Prob. 11. Make use of the results of Prob. 11 to show that

$$\frac{D}{D_{11}} = \frac{K(d_1, \cdots d_n)}{K(d_2, \cdots d_n)} = d_1 + \frac{1|}{|d_2} + \frac{1|}{|d_3} + \frac{1|}{|d_4} + \cdots$$

which is known as a continued fraction.

13. Show that

$$K(d_1, \cdots d_n) = K(d_n, \cdots d_1)$$

and that the recursion formula given in Prob. 11 may alternatively be written

$$K(d_1, \cdots d_n) = d_1 K(d_2, \cdots d_n) + K(d_3, \cdots d_n)$$

14. Show that the partial derivative of a determinant with respect to one of its elements equals the cofactor of that element. In symbols: $\partial A/\partial a_{sk} = A_{sk}$.

15. Consider the cofactors A_{sk} and A_{sr} corresponding to the elements a_{sk} and a_{sr} in the same row of a determinant A. Show that the sum $(A_{sk} + A_{sr})$ is equal to $(-1)^{k-r+1}$ times A_{sk} with the column involving the elements a_{sr} replaced by one with elements $(a_{sr} + a_{sk})$, or to $(-1)^{k-r+1}$ times A_{sr} with the column involving the elements A_{sk} replaced by one in which the elements are $(a_{sk} + a_{sr})$.

16. Using the type of reasoning involved in the previous problem, show that the fourth-order determinant may be written as the following sum of two third-order determinants:

$$\begin{vmatrix} (a_{11}a_{22} - a_{12}a_{21}) & a_{23} & a_{24} \\ (a_{11}a_{32} - a_{12}a_{31}) & a_{33} & a_{34} \\ (a_{11}a_{42} - a_{12}a_{41}) & a_{43} & a_{44} \end{vmatrix} + \begin{vmatrix} a_{21} & a_{22} & (a_{13}a_{24} - a_{14}a_{23}) \\ a_{31} & a_{32} & (a_{13}a_{34} - a_{14}a_{33}) \\ a_{41} & a_{42} & (a_{13}a_{44} - a_{14}a_{43}) \end{vmatrix}$$

or as a variety of obvious modifications of these forms.

17. Express in determinant form the condition that the three straight lines defined by

$$a_{11}x + a_{12}y + a_{13} = 0$$

$$a_{21}x + a_{22}y + a_{23} = 0$$

$$a_{31}x + a_{32}y + a_{33} = 0$$

shall intersect at a common point.

18. In the XY-plane the origin O, and the two points $P(x_1,y_1)$ and $Q(x_2,y_2)$ determine a triangle. Show that the area of this triangle is expressible by means of the determinant

$$\frac{1}{2}\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix}$$

19. Using the result of the previous problem show that the area of a triangle determined by the points (x_1,y_1), (x_2,y_2) and (x_3,y_3) is expressible as

$$\frac{1}{2}\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}$$

Write the condition for which these three points lie on the same straight line.

20. $ax + by + cz + d = 0$ is the equation of a plane. Its intercepts on the coordinate axes are: $x = -d/a$, $y = -d/b$, $z = -d/c$. Let n denote the normal from the origin to the plane. Its direction cosines are:

$$\cos(n,x) = \frac{n}{-\left(\dfrac{d}{a}\right)} \qquad \cos(n,y) = \frac{n}{-\left(\dfrac{d}{b}\right)} \qquad \cos(n,z) = \frac{n}{-\left(\dfrac{d}{c}\right)}$$

Since the sum of the squares of these cosines equals unity, one has

$$n = \frac{d}{\sqrt{a^2 + b^2 + c^2}}$$

Consider a point $P_0(x_0,y_0,z_0)$ for which $ax_0 + by_0 + cz_0 + d = D_0$. Subtracting this equation from the original one gives $a(x - x_0) + b(y - y_0) + c(z - z_0) + D_0 = 0$, from which it is clear that the length of the normal dropped from the point P_0 to the plane is

$$n_0 = \frac{D_0}{\sqrt{a^2 + b^2 + c^2}}$$

These results and that of the previous problem are to be made use of to show: (a) That the equation of a plane passing through the three points $P_1(x_1,y_1,z_1)$, $P_2(x_2,y_2,z_2)$ and $P_3(x_3,y_3,z_3)$ may be written in the form

$$D(x,y,z) = \begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0$$

(b) That the cofactors of the first three elements of the first row, that is,

$$\begin{vmatrix} y_1 & z_1 & 1 \\ y_2 & z_2 & 1 \\ y_3 & z_3 & 1 \end{vmatrix}, \quad \begin{vmatrix} z_1 & x_1 & 1 \\ z_2 & x_2 & 1 \\ z_3 & x_3 & 1 \end{vmatrix}, \quad \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}$$

are equal to the projections of the area of the triangle $P_1P_2P_3$ upon the X-, Y-, Z-axes respectively and that the square root of the sum of the squares of these cofactors equals the area of this triangle. (c) That $D_0 = D(x_0,y_0,z_0)$ divided by this square root equals the normal distance of a point $P_0(x_0,y_0,z_0)$ from the plane, and hence that the volume of the tetrahedron whose vertexes are the points $P_0P_1P_2P_3$ is equal to one-sixth the value of the determinant D_0.

21. Three planes passing through the origin are represented by the equations

$$a_1x + a_2y + a_3z = 0$$

$$b_1x + b_2y + b_3z = 0$$

$$c_1x + c_2y + c_3z = 0$$

Express in determinant form the condition for which these planes intersect in the same straight line and find the expressions for the direction cosines of this line. Given any two planes, what are the direction cosines of their intersection?

22. Write the following determinant as a polynomial in λ:

$$P(\lambda) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix}$$

and obtain expressions for the coefficients of the polynomial in terms of the determinant A and its cofactors. Indicate the forms of these expressions for an nth degree polynomial.

23. Prove that

(a) $$\begin{vmatrix} (a_{11} + b_{11}) & c_{12} \cdots c_{1n} \\ (a_{21} + b_{21}) & c_{22} \cdots c_{2n} \\ \cdots\cdots & \cdots\cdots \\ (a_{n1} + b_{n1}) & c_{n2} \cdots c_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & c_{12} \cdots c_{1n} \\ a_{21} & c_{22} \cdots c_{2n} \\ \cdots & \cdots\cdots \\ a_{n1} & c_{n2} \cdots c_{nn} \end{vmatrix} + \begin{vmatrix} b_{11} & c_{12} \cdots c_{1n} \\ b_{21} & c_{22} \cdots c_{2n} \\ \cdots & \cdots\cdots \\ b_{n1} & c_{n2} \cdots c_{nn} \end{vmatrix}$$

(b) If $\quad c_{ik} = a_{ik} + jb_{ik} \qquad i, k = 1, 2, 3 \qquad j = \sqrt{-1}$

then

$$|c_{ik}| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} - \begin{vmatrix} a_{11} & b_{12} & b_{13} \\ a_{21} & b_{22} & b_{23} \\ a_{31} & b_{32} & b_{33} \end{vmatrix} - \begin{vmatrix} b_{11} & a_{12} & b_{13} \\ b_{21} & a_{22} & b_{23} \\ b_{31} & a_{32} & b_{33} \end{vmatrix} - \begin{vmatrix} b_{11} & b_{12} & a_{13} \\ b_{21} & b_{22} & a_{23} \\ b_{31} & b_{32} & a_{33} \end{vmatrix}$$

$$+j \left\{ \begin{vmatrix} a_{11} & a_{12} & b_{13} \\ a_{21} & a_{22} & b_{23} \\ a_{31} & a_{32} & b_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & b_{12} & a_{13} \\ a_{21} & b_{22} & a_{23} \\ a_{31} & b_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} b_{11} & a_{12} & a_{13} \\ b_{21} & a_{22} & a_{23} \\ b_{31} & a_{32} & a_{33} \end{vmatrix} - \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix} \right\}$$

(c) If $$d_{ik} = a_{ik} - jb_{ik} \qquad i, k = 1, 2, 3$$

then

$$|d_{ik}| = \overline{|c_{ik}|} \qquad \text{complex conjugate of } |c_{ik}|$$

24. Let

$$c_{ik} = a_{ik} + jb_{ik} \qquad i, k = 1, 2, \cdots n \qquad j = \sqrt{-1}$$

Prove that $|c_{ik}|$ is a complex number with the following law of formation:

(a) $|c_{ik}|$ is equal to the sum of 2^n determinants of order n, 2^{n-1} being real and 2^{n-1} pure imaginary.

(b) The real determinants have an even, or zero, number of "b" columns. The imaginary ones have an odd number of "b" columns.

(c) If m is the number of "b" columns in any determinant in the expansion, the sign and complex character of the determinant are given by j^m. For a given value of m, there are several determinants which contain m "b" columns; their number is given by

$$\frac{n(n-1)(n-2)\cdots(n-m+1)}{m\,!}$$

(d) By using the above properties show incidentally that

$$1 + n + \frac{n(n-1)}{2\,!} + \frac{n(n-1)(n-2)}{3\,!} + \cdots + \frac{n\,!}{n\,!} = 2^n$$

25. If $y_{ik} = y_{ik}(t)$, for $i, k = 1, 2, \cdots n$, are n^2 single-valued, differentiable functions of the independent variable t, show that

$$\frac{d}{dt}|y_{ik}| = \begin{vmatrix} y'_{11} & y_{12} & \cdots & y_{1n} \\ y'_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y'_{n1} & y_{n2} & \cdots & y_{nn} \end{vmatrix} + \begin{vmatrix} y_{11} & y'_{12} & \cdots & y_{1n} \\ y_{21} & y'_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y'_{n2} & \cdots & y_{nn} \end{vmatrix} + \cdots + \begin{vmatrix} y_{11} & y_{12} & \cdots & y'_{1n} \\ y_{21} & y_{22} & \cdots & y'_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y'_{nn} \end{vmatrix}$$

$$= \begin{vmatrix} y'_{11} & y'_{12} & \cdots & y'_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{vmatrix} + \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y'_{21} & y'_{22} & \cdots & y'_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{vmatrix} + \cdots + \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y'_{n1} & y'_{n2} & \cdots & y'_{nn} \end{vmatrix}$$

in which

$$y'_{ik} = \frac{dy_{ik}}{dt}$$

26. In terms of the n independent functions

$$y_k = y_k(t) \quad \text{for } k = 1, 2, \cdots n$$

(differentiable up to and including the nth order), construct the following determinant (the so-called Wronskian of those functions):

$$\Delta = \begin{vmatrix} y_1 & y_2 & y_3 & \cdots & y_n \\ y'_1 & y'_2 & y'_3 & \cdots & y'_n \\ y''_1 & y''_2 & y''_3 & \cdots & y''_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_1^{(n-1)} & y_2^{(n-1)} & y_3^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix} \qquad y_k^{(h)} = \frac{d^h}{dt^h}(y_k)$$

If these functions are connected by a linear relation of the form

$$A_1y_1 + A_2y_2 + A_3y_3 + \cdots + A_ny_n = 0$$

in which the A's are constants, show that the above determinant vanishes identically.

Hint. Differentiate the linear relation successively $n - 1$ times so as to obtain

$$\begin{aligned} A_1y'_1 + \cdots + A_ny'_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots & \\ A_1y_1^{(n-1)} + \cdots + A_ny_n^{(n-1)} &= 0 \end{aligned}$$

Together with the original relation, one then has a set of n equations. From these, the value of any function y_k, for example, y_1, and its first $(n - 1)$ derivatives can be obtained. Substitution into the Wronskian, followed by an expansion according to columns, leads to the desired result.

27. Using the determinant, Δ, as defined in Prob. 26, show that

$$\frac{d}{dt}\Delta = \begin{vmatrix} y_1 & y_2 & y_3 & \cdots & y_n \\ y'_1 & y'_2 & y'_3 & \cdots & y'_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_1^{(n-2)} & y_2^{(n-2)} & y_3^{(n-2)} & \cdots & y_n^{(n-2)} \\ y_1^{(n)} & y_2^{(n)} & y_3^{(n)} & \cdots & y_n^{(n)} \end{vmatrix}$$

Hint. Use the result of Prob. 25 and observe the resulting structure of the rows.

28. If (with reference to the situation given in Prob. 27) there exists a set of n relationships (differential equations) of the form

$$p_0y_r^{(n)} + p_1y_r^{(n-1)} + p_2y_r^{(n-2)} + \cdots + p_ny_r = 0$$

for

$$r = 1, 2, \cdots n$$

in which the coefficients $p_0, p_1, p_2, \cdots p_n$ are constant or variable, show that

(a)
$$\frac{d}{dt}\Delta = -\frac{p_1}{p_0}\Delta$$

(b)
$$\Delta = \Delta_0 e^{-\int (p_1/p_0)\,dt}$$

in which Δ_0 is the integration constant.

Hint. Give r the values $1, 2, \cdots n$ and obtain, from each equation, the value of $y_r^{(n)}$. Substitute these values into the last row of the expression in Prob. 27.

29. Given

$x_k = x_k(t)$ for $k = 1, 2, \cdots n$ (a system of n unknown functions of t)

$y_k = y_k(t)$ for $k = 1, 2, \cdots n$ (a system of n known functions of t)

and

a_{ik} for $i, k = 1, 2, \cdots n$ (a collection of n^2 constants)

These quantities are related by the following system of first-order, first-degree, linear equations

$$a_{1k}\dot{x}_1 + a_{2k}\dot{x}_2 + \cdots + a_{nk}\dot{x}_n = y_k \quad \text{for } k = 1, 2, \cdots n$$

Show that a solution of this system (the particular one) is given by

$$x_k = \frac{A_{1k}\int y_1\,dt + A_{2k}\int y_2\,dt + \cdots + A_{nk}\int y_n\,dt}{|a_{ik}|}$$

in which the A_{ik}'s are cofactors of the determinant $|a_{ik}|$.

30. Show that

$$\begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2^2 & \cdots & 2^{n-1} \\ 1 & 3 & 3^2 & \cdots & 3^{n-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & n & n^2 & \cdots & n^{n-1} \end{vmatrix} = 1\,! \times 2\,! \times 3\,! \cdots (n-1)\,!$$

Hint. Reduce the determinant to the diagonal form. Use Barlow's tables of squares (pages 202 to 206) for the powers of integer numbers and observe the law of formation.

31.

$$u_k = u_k(x_1 \cdots x_n) \quad \text{for } k = 1, 2, \cdots n$$

are n single-valued differentiable functions of the independent variables $x_1, x_2, \cdots x_n$.

The "Jacobian" of these functions is, by definition, the following functional determinant:

$$J = \left(\frac{u_1 u_2 \cdots u_n}{x_1 x_2 \cdots x_n}\right) = \begin{vmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} & \cdots & \dfrac{\partial u_1}{\partial x_n} \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} & \cdots & \dfrac{\partial u_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial u_n}{\partial x_1} & \dfrac{\partial u_n}{\partial x_2} & \cdots & \dfrac{\partial u_n}{\partial x_n} \end{vmatrix}$$

Suppose the variables $x_1 \cdots x_n$ are changed to the new independent variables $z_1 \cdots z_n$ according to the equations of transformation

$$x_k = x_k(z_1 \cdots z_n) \quad \text{for } k = 1, 2, \cdots n$$

The original u_k functions are now

$$u_k = u_k(z_1 \cdots z_n) \quad \text{for } k = 1, 2, \cdots n$$

and their Jacobian with respect to the variables $z_1 \cdots z_n$, for example, J_1, differs from J only in that the variables x are replaced by corresponding z's.

(a) Show that the above Jacobians are connected by the relation

$$J_1 = \begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} & \cdots & \frac{\partial x_1}{\partial z_n} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} & \cdots & \frac{\partial x_2}{\partial z_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_n}{\partial z_1} & \frac{\partial x_n}{\partial z_2} & \cdots & \frac{\partial x_n}{\partial z} \end{vmatrix} \times J$$

(b) Extend the above result so as to consider subsequent transformations of the form

$$z_j = z_j(r_1, r_2, \cdots r_n) \quad \text{for } j = 1\ 2, \cdots n$$

$$r_p = r_p(t_1, t_2, \cdots t_n) \quad \text{for } p = 1, 2, \cdots n$$

(c) What happens with the last Jacobian when any intermediate functional determinant is identically zero? *Hint*. Apply the rule for differentiation which reads:

$$\frac{\partial u_r}{\partial z_s} = \sum_{i=1}^{i=n} \frac{\partial u_r}{\partial x_i} \frac{\partial x_i}{\partial z_s} \qquad r, s = 1, 2, \cdots n$$

32. Let

$$a_{jk} = a_{jk}(x_1, x_2, \cdots x_n) \quad \text{for } j, k = 1, 2, \cdots n$$

be a system of n^2 differentiable functions of the independent variables $x_1, x_2 \cdots x_n$. Through the introduction of a new set of independent variables $\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n$ by means of the functional relations

$$x_k = x_k(\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n) \quad \text{for } k = 1, 2, \cdots n$$

the system of functions a_{jk} in the old variables goes over into the transformed system $\bar{a}_{jk}$ in the new variables.

Accepting the result that a_{jk} goes over into $\bar{a}_{jk}$ in accordance with the law of transformation

$$\bar{a}_{pq} = \sum_{j=1}^{j=n} \sum_{k=1}^{k=n} a_{jk} \frac{\partial x_j}{\partial x_p} \frac{\partial x_k}{\partial x_q} \qquad (p, q = 1, 2, \cdots n)$$

prove that the determinant $|a_{jk}|$ is transformed as indicated by

$$\overline{|a_{pq}|} = |a_{jk}| \times J^2$$

in which

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial \bar{x}_1} & \frac{\partial x_1}{\partial \bar{x}_2} & \cdots & \frac{\partial x_1}{\partial \bar{x}_n} \\ \frac{\partial x_2}{\partial \bar{x}_1} & \frac{\partial x_2}{\partial \bar{x}_2} & \cdots & \frac{\partial x_2}{\partial \bar{x}_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_n}{\partial \bar{x}_1} & \frac{\partial x_n}{\partial \bar{x}_2} & \cdots & \frac{\partial x_n}{\partial \bar{x}_n} \end{vmatrix}$$

33. The expression for the three-dimensional volume element in a general system of co-ordinates is given by

$$dV = \sqrt{|g_{jk}|}\, d\bar{x}_1\, d\bar{x}_2\, d\bar{x}_3$$

in which

$$g_{jk} = \sum_{\nu=1}^{\nu=n} \frac{\partial x_\nu}{\partial \bar{x}_j} \frac{\partial x_\nu}{\partial \bar{x}_k} \qquad \nu = 1, 2, \cdots n \qquad j = 1, 2, \cdots n \qquad k = 1, 2, \cdots n$$

If the co-ordinate system is orthogonal, the g_{jk} system has the property

$$g_{jk} \begin{cases} \neq 0 & \text{for } j = k \\ = 0 & \text{for } j \neq k \end{cases}$$

Check the values of $\sqrt{|g_{jk}|}$ for the different co-ordinate systems and laws of co-ordinate transformation given in the following table:

Name	Equations of Transformation	$\sqrt{\lvert g_{jk}\rvert}$
Cartesian	$x_1 = \bar{x}_1$ $x_2 = \bar{x}_2$ $x_3 = \bar{x}_3$	1
Circular cylindrical	$x_1 = \bar{x}_1 \cos \bar{x}_2$ $x_2 = \bar{x}_1 \sin \bar{x}_2$ $x_3 = \bar{x}_3$	$\bar{x}_1$
Elliptic cylindrical	$x_1 = c \cosh \bar{x}_1 \cos \bar{x}_2$ $x_2 = c \sinh \bar{x}_1 \sin \bar{x}_2$ $x_3 = \bar{x}_3;\ c = \text{const.}$	$c^2(\cosh^2 \bar{x}_1 - \cos^2 \bar{x}_2)$
Parabolic cylindrical	$x_1 = \frac{1}{2}(\bar{x}_1^2 - \bar{x}_2^2)$ $x_2 = \bar{x}_1\bar{x}_2$ $x_3 = \bar{x}_3$	$(\bar{x}_1^2 + \bar{x}_2^2)$
Bipolar cylindrical	$x_1 = \dfrac{a \sinh \bar{x}_1}{\cosh \bar{x}_1 - \cos \bar{x}_2}$ $x_2 = \dfrac{a \sin \bar{x}_2}{\cosh \bar{x}_1 - \cos \bar{x}_2}$ $x_3 = \bar{x}_3;\ a = \text{const.}$	$\dfrac{a^2}{(\cosh \bar{x}_1 - \cos \bar{x}_2)^2}$
Spheroidal	$x_1 = c \cosh \bar{x}_1 \cos \bar{x}_2;\ c = \text{const.}$ $x_2 = c \sinh \bar{x}_1 \sin \bar{x}_2 \cos \bar{x}_3$ $x_3 = c \sinh \bar{x}_1 \sin \bar{x}_2 \sin \bar{x}_3$	$c^3(\cosh^2 \bar{x}_1 - \cos^2 \bar{x}_2) \sinh \bar{x}_1 \sin \bar{x}_2$
Spherical	$\bar{x}_1 = \bar{x}_1 \cos \bar{x}_2 \sin \bar{x}_3$ $\bar{x}_2 = \bar{x}_1 \sin \bar{x}_2 \sin \bar{x}_3$ $\bar{x}_3 = \bar{x}_1 \cos \bar{x}_3$	$\bar{x}_1^2 \sin \bar{x}_3$

34. Given the multiple integral

$$I = \int\int \cdots \int F(x_1, x_2, \cdots x_n)\, dx_1\, dx_2 \cdots dx_n$$

in which $x_1, x_2, \cdots x_n$ are the independent variables. If new variables $\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n$

are introduced in accordance with the relations

$$x_k = x_k(\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n) \quad \text{for } k = 1, 2, \cdots n$$

it can be shown that the above integral becomes

$$I = \int\int \cdots \int JF(\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n)\, d\bar{x}_1\, d\bar{x}_2 \cdots d\bar{x}_n$$

in which J is the determinant given in Prob. 32.

Compute the value of the determinant J for each set of transformation functions given in the second column of the table in Prob. 33.

35. Given the following system of $m + n$ linear equations involving the $m + n$ unknowns x_λ for $\gamma = 1, 2, \cdots n$, and e_ρ for $\rho = 1, 2, \cdots m$, with $n > m$:

$$\sum_{\mu=1}^{\mu=n} A_{\lambda\mu}x_\mu - ex_\lambda + \sum_{\rho=1}^{\rho=m} e_\rho a_{\rho\lambda} = 0 \quad \text{for } \lambda = 1, 2, \cdots n$$

$$\sum_{\mu=1}^{\mu=n} a_{\rho\mu}x_\mu = 0 \quad \text{for } \rho = 1, 2, \cdots m$$

(a) Write in determinant form, according to the above order of these equations, the condition for the existence of nontrivial solutions.

(b) Show that this determinant is a polynomial in e of the degree $n - m$.

(c) Using Laplace's development with respect to the last m rows, show that the total number of mth-order minors which can be formed is given by

$$\frac{(n + m)(n + m - 1)(n + m - 2) \cdots (n + 1)}{m\,!} \qquad (m < n)$$

and that only

$$\frac{n(n - 1)(n - 2) \cdots (n - m + 1)}{m\,!}$$

of these are not necessarily zero (the rest being identically zero).

CHAPTER II

Matrices

1. Linear transformations

In Art. 8 of Ch. I, the system of linear equations

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned}\right\} \qquad [1]$$

is considered to relate a set of unknown quantities $x_1 \cdots x_n$ to a set of known quantities $y_1 \cdots y_n$. An alternate point of view, which plays an important part in the analysis of physical problems, is to look upon these equations as relating a set of given variables $x_1 \cdots x_n$ to a new set of variables $y_1 \cdots y_n$. The equations are said to *transform* the old variables into new ones. In this sense the set of Eqs. 1 is spoken of as a *linear transformation.*

The transformation is characterized completely by the coefficients a_{ik}. Since not only the values of the coefficients but also their relative positions in the equations are significant in this respect, a symbolic form for the characterization of the linear transformation is given by means of the rectangular array

$$\mathcal{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \qquad [2]$$

which is called the *matrix* of the transformation.

Offhand, it might be thought that the determinant could serve this purpose, but here it must be recalled that the determinant is a *function* of the coefficients and not merely a symbolic representation of them in their relative orientations. The determinant has a value; the matrix, on the other hand, is merely a picture and has no value other than that which it conveys by its structural composition. For the moment it has but one reason for existence, namely, that it is easier to write down than the system of Eqs. 1, yet contains essentially the same information.

In outward appearance the determinant differs from the matrix only in that the latter is enclosed in square brackets whereas in the determinant the elements are enclosed between vertical lines. In this book, an upper-case script letter is used to denote the matrix and the upper-case

italic letter denotes the corresponding determinant. Thus the matrix 2 is denoted by $\mathfrak{A}$ and the corresponding determinant by A.

Just as determinants facilitate the study of linear equations and related problems, so the algebra of matrices is justified in that it facilitates the manipulation of several sets of linear transformations. The rules of matrix algebra are chosen with this end in view and are otherwise quite arbitrary except that they must, of course, be self-consistent.

A matrix does not necessarily have the same numbers of rows as columns. Its form is therefore more flexible than that of a determinant, which must always be square. If the matrix has m rows and n columns, it is referred to as having the form m by n, or it may simply be called an (mn)-matrix. When m equals n, the matrix is said to be of order n. In the extreme cases in which the matrix consists merely of a single row or a single column, it is referred to as a *row matrix* or a *column matrix* respectively. These special forms occur frequently enough to warrant the use of some abbreviated notation to distinguish them.

The notation used in this text for a row matrix is

$$\underline{\alpha} = [\alpha_1 \quad \alpha_2 \cdots \alpha_n] \tag{3}$$

A column matrix is indicated as shown by the equation

$$\beta] = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix} \tag{4}$$

At times it becomes necessary to have the abbreviated symbol for a matrix indicate the number of rows and columns involved. A matrix having m rows and n columns is written

$$[a_{mn}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{5}$$

A less specific notation, however, is sufficient when the number of rows and columns either is clear from the discussion involved or is immaterial.

If a rule for the multiplication of matrices is properly chosen, the linear transformation, Eq. 1, may be written as a matrix equation. One of the four alternative rules for forming the product of two determinants (discussed in Art. 7, Ch. I) may be used for this purpose. It is conventional to choose the rule expressed by Eq. 34, Ch. I. Thus in order to form the product $\mathfrak{C}$ of two matrices $\mathfrak{A}$ and $\mathfrak{B}$, as indicated by the equation

$$\mathfrak{A} \times \mathfrak{B} = \mathfrak{C} \tag{6}$$

the elements in $\mathcal{C}$ are obtained by multiplying the rows of $\mathcal{A}$ by the columns of $\mathcal{B}$, as described in detail for the multiplication of determinants.

With the adoption of this convention, the matrix equation replacing Eq. 1 reads

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad [7]$$

which may be abbreviated as

$$\mathcal{A} \times x] = y] \qquad [8]$$

The element y_1, which is the (1,1)-element in the column matrix $y]$, is obtained by multiplying the elements in the *first* row of $\mathcal{A}$ by the corresponding elements of the *first* column (the only column) of $x]$ and adding the results. This gives

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \qquad [9]$$

which is the first equation in the set 1. Similarly, y_2 which is the (2,1)-element in the matrix $y]$, is formed from the *second* row of $\mathcal{A}$ and the *first* column of $x]$, thus

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \qquad [10]$$

This is the second equation in the set 1. The remaining equations in this set are similarly contained in the matrix equation 7 or 8.

It should be noted that the column matrix $y]$ has no elements for the position (1, 2), (1, 3) $\cdots$ or (2, 2), (2, 3) $\cdots$, etc., because it has only one column — the first column. For the missing elements, or zero elements in $y]$, the matrix product expressed by Eq. 7 automatically yields zeros according to the adopted product-forming rule because the column matrix $x]$ has no second, third, $\cdots$, etc., columns; that is, the elements in these missing columns of the matrix $x]$ are all zero.

The present argument requires for completeness a statement of the rule that two matrices are equal only when all corresponding elements are equal. Specifically, if two matrices $\mathcal{P}$ and $\mathcal{Q}$ with elements p_{rs} and q_{rs} are to be equal, that is, if

$$\mathcal{P} = \mathcal{Q} \qquad [11]$$

it is necessary that

$$p_{rs} = q_{rs} \qquad [12]$$

for all values of the indexes r and s. For example, the left-hand side of Eq. 9 is the (1,1)-element of the product of $\mathcal{A}$ and $x]$ in Eq. 7, and y_1,

the right-hand side of Eq. 9, is the (1,1)-element of the matrix $y]$ in Eq. 7. This rule for the equality of two matrices is, however, so obvious that it hardly needs emphasis.

It should be clearly recognized that the objective of obtaining a matrix equation equivalent to the transformation 1 could have been met just as well by the adoption of any one of the other three methods of forming elements in the product matrix according to the rules of multiplying determinants and by suitable modification of the form of the matrix equation. The choice of the rule by which the rows of the first matrix in the product are multiplied by the columns of the second is arbitrary, but once it is adopted, consistency requires that it be looked upon as *the* rule. There are no alternatives as in the multiplication of determinants, because the alternative procedures, if applied to the example illustrated by Eq. 7, do not all yield the same final set of equations.

That the adopted rule for matrix multiplication is not compatible with the commutative law in forming algebraic products is likewise apparent. Thus if $\mathcal{A}$ and $\mathcal{B}$ are two matrices, the product $\mathcal{A} \times \mathcal{B}$ is different (and has a different manipulative significance) from the product $\mathcal{B} \times \mathcal{A}$. Thus in general

$$\mathcal{A} \times \mathcal{B} \neq \mathcal{B} \times \mathcal{A} \tag{13}$$

in matrix multiplication.

The utility of being able to write a transformation in the compact matrix form is appreciated when several transformations are to be carried out in tandem. In addition to the transformation 1 from the variables $x_1 \cdots x_n$ to the variables $y_1 \cdots y_n$, it may be necessary to transform subsequently to a set of variables $z_1 \cdots z_n$ by means of the equations

$$\left.\begin{array}{l} b_{11}y_1 + b_{12}y_2 + \cdots + b_{1n}y_n = z_1 \\ b_{21}y_1 + b_{22}y_2 + \cdots + b_{2n}y_n = z_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{n1}y_1 + b_{n2}y_2 + \cdots + b_{nn}y_n = z_n \end{array}\right\} \tag{14}$$

with the transformation matrix

$$\mathcal{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \tag{15}$$

With the additional definition of the column matrix

$$z] = \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{bmatrix} \tag{16}$$

this transformation takes the compact form

$$\mathcal{B} \times y] = z] \tag{17}$$

The direct transformation from the variables $x_1 \cdots x_n$ to the final variables $z_1 \cdots z_n$ is obtained by substitution of the expression for $y]$ according to Eq. 8 into Eq. 17, which gives

$$\mathcal{B} \times \mathcal{A} \times x] = z] \tag{18}$$

or, more compactly,

$$\mathcal{C} \times x] = z] \tag{19}$$

in which the resultant transformation matrix $\mathcal{C}$ is given by the product

$$\mathcal{C} = \mathcal{B} \times \mathcal{A} \tag{20}$$

The convenience of the matrix method of handling transformations becomes clear if it is contrasted with the additional labor involved when a specific example of the above transformation is carried through by means of the usual substitution process. The individual transformations may be assumed to be the following simple ones:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= y_1 \\ a_{21}x_1 + a_{22}x_2 &= y_2 \end{aligned}\right\} \tag{21}$$

and

$$\left.\begin{aligned} b_{11}y_1 + b_{12}y_2 &= z_1 \\ b_{21}y_1 + b_{22}y_2 &= z_2 \end{aligned}\right\} \tag{22}$$

Substitution of Eqs. 21 into Eqs. 22 gives

$$\left.\begin{aligned} b_{11}(a_{11}x_1 + a_{12}x_2) + b_{12}(a_{21}x_1 + a_{22}x_2) &= z_1 \\ b_{21}(a_{11}x_1 + a_{12}x_2) + b_{22}(a_{21}x_1 + a_{22}x_2) &= z_2 \end{aligned}\right\} \tag{23}$$

Multiplying out and collecting terms with x_1 and x_2, one obtains

$$\left.\begin{aligned} (b_{11}a_{11} + b_{12}a_{21})x_1 + (b_{11}a_{12} + b_{12}a_{22})x_2 &= z_1 \\ (b_{21}a_{11} + b_{22}a_{21})x_1 + (b_{21}a_{12} + b_{22}a_{22})x_2 &= z_2 \end{aligned}\right\} \tag{24}$$

The coefficients in these equations are recognized as the elements of a matrix formed from the product

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \times \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{25}$$

The ease of manipulation and circumspection resulting from the use of the corresponding matrix equations is proved even more conclusively in the applications discussed in Ch. IV. Meanwhile it is necessary to study the rules governing various additional algebraic manipulations with matrices.

2. Addition of matrices

By the sum of two matrices $\mathcal{A}$ and $\mathcal{B}$ is meant a resultant matrix $\mathcal{C}$ whose elements are equal to the sums of corresponding elements of $\mathcal{A}$ and $\mathcal{B}$. Explicitly, if

$$\mathcal{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad \mathcal{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} \tag{26}$$

then

$$\mathcal{A} + \mathcal{B} = \mathcal{C} = \begin{bmatrix} (a_{11} + b_{11}) & \cdots & (a_{1n} + b_{1n}) \\ \cdots & \cdots & \cdots \\ (a_{m1} + b_{m1}) & \cdots & (a_{mn} + b_{mn}) \end{bmatrix} \tag{27}$$

Similarly, the difference $\mathcal{A} - \mathcal{B}$ yields a matrix whose elements are the differences of corresponding elements of $\mathcal{A}$ and $\mathcal{B}$.

Evidently the matrices $\mathcal{A}$ and $\mathcal{B}$ should have the same number of rows and columns, or the missing rows and columns of one may be regarded as composed entirely of zeros.

3. Multiplication by a factor

If a matrix is multiplied by a factor, each element of the matrix becomes multiplied by this factor; thus

$$k \times \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} ka_{11} & \cdots & ka_{1n} \\ \cdots & \cdots & \cdots \\ ka_{m1} & \cdots & ka_{mn} \end{bmatrix} \tag{28}$$

As is clear from the fact that if the factor k is zero, the matrix must be zero, which is true only when all elements of the matrix are zero.

It should be observed that this rule is distinctly different from the corresponding one for determinants, in that the determinant becomes multiplied by a factor when the elements of only one row or column are multiplied by this factor.

4. Multiplication of matrices

The operation of multiplication is introduced in Art. 1 above, but several additional remarks are necessary to complete the discussion. The individual matrices entering into the product $\mathcal{A} \times \mathcal{B}$ may not have the same number of rows as columns, but it is clear from the manner in which this product is carried out that the number of columns in $\mathcal{A}$ (the number of elements in a row of $\mathcal{A}$) must equal the number of rows in $\mathcal{B}$ (the number of elements in a column of $\mathcal{B}$); otherwise some of the

products of elements in the rows of $\mathfrak{A}$ with corresponding elements in the columns of $\mathfrak{B}$ cannot be completed, for want of either one or the other of the corresponding elements. For example, when expansion of the following product,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \qquad [29]$$

is attempted, it is found that there are no elements in the matrix $\mathfrak{B}$ by which the elements a_{13} and a_{23} may be multiplied. These matrices are said to be *nonconformable*. They become *conformable* when a third row is added to the matrix $\mathfrak{B}$. This row may, of course, consist entirely of zeros.

Curiously enough, conformability does not require that the number of rows in $\mathfrak{A}$ be equal to the number of columns in $\mathfrak{B}$. Thus the formation of the product

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \times \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} \qquad [30]$$

is considered straightforward, it being tacitly understood that the matrix $\mathfrak{B}$ is interpreted as equivalent to

$$\begin{bmatrix} b_{11} & 0 \\ b_{21} & 0 \\ b_{31} & 0 \end{bmatrix} \qquad [31]$$

but *never* that it is equivalent to

$$\begin{bmatrix} 0 & b_{11} \\ 0 & b_{21} \\ 0 & b_{31} \end{bmatrix} \qquad [32]$$

It is useful to observe that the matrix formed from the product $\mathfrak{A} \times \mathfrak{B}$ has as many rows as $\mathfrak{A}$ and as many columns as $\mathfrak{B}$. For example, the product 30 yields a matrix with two rows and one column. In general, if the matrices are denoted more specifically as $[a_{mn}]$ and $[b_{np}]$, then

$$[a_{mn}] \times [b_{np}] = [c_{mp}] \qquad [33]$$

These matrices are conformable because $\mathfrak{A}$ has n columns and $\mathfrak{B}$ has n rows.

Similarly the product of three conformable matrices reads

$$[a_{mn}] \times [b_{np}] \times [c_{pq}] = [d_{mq}] \qquad [34]$$

Here the resultant matrix has as many rows as the first matrix and as many columns as the last one entering into the triple product.

To carry out the product of three or more matrices such as $\mathfrak{A} \times \mathfrak{B} \times \mathfrak{C}$, the component product $\mathfrak{A} \times \mathfrak{B}$ may be evaluated first and the resultant

matrix multiplied into $\mathcal{C}$, or $\mathcal{A}$ may be multiplied by the resultant of $\mathcal{B} \times \mathcal{C}$. In symbols this procedure is stated by the equation

$$\mathcal{A} \times \mathcal{B} \times \mathcal{C} = (\mathcal{A} \times \mathcal{B}) \times \mathcal{C} = \mathcal{A} \times (\mathcal{B} \times \mathcal{C}) \qquad [35]$$

In other words, the *associative* law holds in matrix multiplication.

This fact may prove useful in minimizing the amount of labor involved in carrying out a multiple product. Observe that the multiplication of the two matrices

$$[a_{mn}] \times [b_{np}] \qquad [36]$$

involves the operations

$$\begin{aligned} &m \times n \times p \quad \text{multiplications} \\ &m \times (n-1) \times p \quad \text{additions} \end{aligned} \qquad [37]$$

Consequently a triple product in the association

$$([a_{mn}] \times [b_{np}]) \times [c_{pq}] \qquad [38]$$

involves the operations

$$\begin{aligned} &m \times n \times p + m \times p \times q \quad \text{multiplications} \\ &m \times (n-1) \times p + m \times (p-1) \times q \quad \text{additions} \end{aligned} \qquad [39]$$

and in the association

$$[a_{mn}] \times ([b_{np}] \times [c_{pq}]) \qquad [40]$$

the operations are

$$\begin{aligned} &n \times p \times q + m \times n \times q \quad \text{multiplications} \\ &n \times (p-1) \times q + m \times (n-1) \times q \quad \text{additions} \end{aligned} \qquad [41]$$

As a numerical illustration let it be supposed that $m = 1$, $n = 6$, $p = 1$, and $q = 4$. Then the operations 39 are ten multiplications and five additions, whereas the operations 41 amount to 48 multiplications and 20 additions.

Since, as pointed out earlier, the commutative law does not hold, the order in which the matrices enter into the multiple product must be carefully observed. In Eq. 35, for example, the component product $(\mathcal{A} \times \mathcal{B})$ is *post*multiplied by $\mathcal{C}$, and the component product $(\mathcal{B} \times \mathcal{C})$ is *pre*multiplied by $\mathcal{A}$.

In this connection it is to be noted that the *distributive* law is valid also, with the caution that attention must again be given to the distinction between premultiplication and postmultiplication. Two cases are illustrated in the equations

$$(\mathcal{A} + \mathcal{B}) \times \mathcal{C} = \mathcal{A} \times \mathcal{C} + \mathcal{B} \times \mathcal{C} \qquad [42]$$

and

$$\mathcal{C} \times (\mathcal{A} + \mathcal{B}) = \mathcal{C} \times \mathcal{A} + \mathcal{C} \times \mathcal{B} \tag{43}$$

It is useful to observe that since the rule for matrix multiplication agrees with that for the multiplication of determinants, the determinant of the multiple product

$$\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \cdots \times \mathcal{G} \tag{44}$$

has the value

$$A \times B \times C \times \cdots \times G \tag{45}$$

5. Some special forms of square matrices

A square matrix of the form

$$\mathcal{D} = \begin{bmatrix} d_{11} & 0 & 0 \cdots 0 \\ 0 & d_{22} & 0 \cdots 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots d_{nn} \end{bmatrix} \tag{46}$$

in which all the elements are zero except those on the principal diagonal, is called a *diagonal* matrix.

If, in addition, all the diagonal elements are unity, the matrix is referred to as the *identity* or *unit* matrix. It is designated in this book by the special letter $\mathcal{U}$ and has the appearance

$$\mathcal{U} = \begin{bmatrix} 1 & 0 & 0 \cdots 0 \\ 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots 1 \end{bmatrix} \tag{47}$$

If the matrix of the transformation 1 is $\mathcal{U}$, the equations read

$$\left. \begin{aligned} x_1 &= y_1 \\ x_2 &= y_2 \\ &\cdots\cdots \\ x_n &= y_n \end{aligned} \right\} \tag{48}$$

This set is referred to as the identity transformation, since the old variables $x_1 \cdots x_n$ and the new variables $y_1 \cdots y_n$ are identical. Any matrix either pre- or postmultiplied by $\mathcal{U}$ is equal to itself.

The unit matrix multiplied by a factor k becomes

$$\mathcal{K} = \begin{bmatrix} k & 0 & 0 \cdots 0 \\ 0 & k & 0 \cdots 0 \\ 0 & 0 & k \cdots 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots k \end{bmatrix} \tag{49}$$

This is called a *scalar* matrix because any matrix either pre- or postmultiplied by $\mathcal{K}$ is simply multiplied by the scalar factor k.

When a matrix is multiplied by the diagonal matrix $\mathcal{D}$, Eq. 46, the results for pre- and postmultiplication differ in the following manner. The formation of $\mathcal{D} \times \mathcal{A}$ has the effect of multiplying the *rows* of $\mathcal{A}$ by $d_{11}, d_{22}, \cdots d_{nn}$ respectively, whereas in the formation of $\mathcal{A} \times \mathcal{D}$ the *columns* of $\mathcal{A}$ are so multiplied.

A square matrix $\mathcal{A}$ whose elements satisfy the conditions

$$a_{ki} = a_{ik} \tag{50}$$

evidently possesses symmetry about its principal diagonal. It is, therefore, called a *symmetrical* matrix. The elements on its principal diagonal are, of course, arbitrary.

On the other hand, if the elements are subjected to the conditions

$$a_{ki} = -a_{ik} \tag{51}$$

those on the principal diagonal must be zero. In this case the matrix is said to be *skew* symmetric.

Interchanging rows and columns has no effect upon a symmetrical matrix, but in the case of a skew symmetrical matrix it amounts to multiplying by the factor -1. It is, therefore, seen that if both $\mathcal{A}$ and $\mathcal{B}$ are symmetrical or both are skew symmetrical, $\mathcal{A} \times \mathcal{B} = \mathcal{B} \times \mathcal{A}$; that is, the matrix product is commutative in these special cases.

6. Inverse, adjoint, transposed, reciprocal, and orthogonal matrices

The rank of a square matrix of order n is equal to that of its determinant. When the rank is n, the determinant is not zero, and the matrix is said to be *nonsingular*. The matrix is singular if its rank is less than n, that is, if its determinant is zero.

When the matrix $\mathcal{A}$ of the transformation 1 is nonsingular the transformation may, according to the discussion of Arts. 8 and 9 of the previous chapter, be inverted to yield

$$\left.\begin{array}{l} b_{11}y_1 + b_{12}y_2 + \cdots + b_{1n}y_n = x_1 \\ b_{21}y_1 + b_{22}y_2 + \cdots + b_{2n}y_n = x_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{n1}y_1 + b_{n2}y_2 + \cdots + b_{nn}y_n = x_n \end{array}\right\} \tag{52}$$

The matrix of this inverse transformation

$$\mathcal{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \tag{53}$$

is referred to as the *inverse* of $\mathcal{A}$. This is written

$$\mathcal{A}^{-1} = \mathcal{B} \quad [54]$$

The matrix $\mathcal{A}$ evidently does not possess an inverse when it is singular. It is clear that all matrices are singular in which the number of rows is not equal to the number of columns, because the system of equations represented by such matrices does not have unique solutions. A square matrix does not possess an inverse if its rank is less than its order.

A few additional remarks are in order regarding the inversion of a linear transformation as viewed from the corresponding matrix equations. Writing the matrix equation for the transformation 1

$$\mathcal{A}x] = y] \quad [55]$$

one is tempted to solve this for the matrix $x]$ by simply dividing the equation through by $\mathcal{A}$. This form of solution is, however, meaningless without an interpretation of the operation of division by a matrix.

Such an interpretation is arrived at if one observes first that the elements of the inverse matrix $\mathcal{B}$, Eq. 53, are given by Eq. 48, Ch. I, which reads

$$b_{rs} = \frac{A_{sr}}{A} \quad [56]$$

in which A_{sr} are the cofactors of the determinant A corresponding to the matrix $\mathcal{A}$. The result expressed by Eq. 55, Ch. I, coupled with a recognition of the similarity between the processes of determinant multiplication and matrix multiplication then shows that

$$\mathcal{A} \times \mathcal{A}^{-1} = \mathcal{A}^{-1} \times \mathcal{A} = \mathcal{U} \quad [57]$$

that is, the product of a nonsingular matrix with its inverse equals the unit matrix, Eq. 47. Hence it becomes clear that the matrix equation 55 is solved by premultiplying on both sides by $\mathcal{A}^{-1}$, which gives

$$\mathcal{U}x] = x] = \mathcal{A}^{-1}y] \quad [58]$$

The operation of division by a matrix is, therefore, interpreted in terms of multiplication by means of the inverse matrix.

Division by a matrix is subject to several restrictions. First of all, the matrix by which a matrix equation is to be divided must be nonsingular. Furthermore, in a matrix equation of the form

$$\mathcal{A} \times \mathcal{B} \times \mathcal{C} = \mathcal{G} \quad [59]$$

it is not possible in a single step to divide out $\mathcal{B}$ so as to get a solution for $\mathcal{A} \times \mathcal{C}$.

To obtain this result, a more lengthy procedure such as the following

must be used. The equation may first be postmultiplied by $\mathcal{C}^{-1}$, giving

$$\mathcal{A} \times \mathcal{B} = \mathcal{G} \times \mathcal{C}^{-1} \tag{60}$$

Next it is postmultiplied by $\mathcal{B}^{-1}$. It then reads

$$\mathcal{A} = \mathcal{G} \times \mathcal{C}^{-1} \times \mathcal{B}^{-1} \tag{61}$$

Finally it is postmultiplied by $\mathcal{C}$, giving the desired result

$$\mathcal{A} \times \mathcal{C} = \mathcal{G} \times \mathcal{C}^{-1} \times \mathcal{B}^{-1} \times \mathcal{C} \tag{62}$$

The same result may alternatively be obtained by first premultiplying Eq. 59 by $\mathcal{A}^{-1}$ so that it becomes

$$\mathcal{B} \times \mathcal{C} = \mathcal{A}^{-1} \times \mathcal{G} \tag{63}$$

then premultiplying by $\mathcal{B}^{-1}$ to obtain

$$\mathcal{C} = \mathcal{B}^{-1} \times \mathcal{A}^{-1} \times \mathcal{G} \tag{64}$$

and finally premultiplying by $\mathcal{A}$ so as to have

$$\mathcal{A} \times \mathcal{C} = \mathcal{A} \times \mathcal{B}^{-1} \times \mathcal{A}^{-1} \times \mathcal{G} \tag{65}$$

It should be observed that obtaining the desired result in this case requires not only that $\mathcal{B}$ be nonsingular, but also that either $\mathcal{A}$ or $\mathcal{C}$ be nonsingular. If this additional condition is not met, the desired result cannot be obtained.

Because of these restrictions on the process of division by a matrix, it is best not to use the term " division " at all. In matrix algebra pre- or postmultiplication by an inverse matrix is possible (if the given matrix is nonsingular), but the operation of division is said in general not to exist.

By means of Eq. 56 the inverse of $\mathcal{A}$ may be written in the form

$$\mathcal{A}^{-1} = \begin{bmatrix} \dfrac{A_{11}}{A} & \dfrac{A_{21}}{A} & \cdots & \dfrac{A_{n1}}{A} \\ \dfrac{A_{12}}{A} & \dfrac{A_{22}}{A} & \cdots & \dfrac{A_{n2}}{A} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{A_{1n}}{A} & \dfrac{A_{2n}}{A} & \cdots & \dfrac{A_{nn}}{A} \end{bmatrix} \tag{66}$$

The process of finding this inverse may be described as follows. First, each element in $\mathcal{A}$ is replaced by the quotient of its corresponding cofactor and the determinant of $\mathcal{A}$. Second, the rows and columns in this resulting matrix are interchanged. This procedure evidently involves a large amount of labor when the order of the given matrix is high.

There are other methods by which the inverse of a matrix may be found. One of these, which is usually considerably shorter than that stated here, is described in the next article. Additional methods are discussed in Art. 11.

When the given matrix is of the diagonal form, Eq. 46, its inverse is also of the diagonal form and is given by

$$\mathfrak{D}^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 & 0 \cdots 0 \\ 0 & d_{22}^{-1} & 0 \cdots 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & \cdots \quad d_{nn}^{-1} \end{bmatrix} \qquad [67]$$

This result follows from the fact that all cofactors except those for the diagonal elements in $\mathfrak{D}$ are zero, and these cofactors are

$$D_{kk} = d_{11}d_{22} \cdots d_{k-1,k-1}d_{k+1,k+1} \cdots d_{nn} \qquad [68]$$

while the determinant of $\mathfrak{D}$ is

$$D = d_{11}d_{22} \cdots d_{nn} \qquad [69]$$

Hence

$$\frac{D_{kk}}{D} = d_{kk}^{-1} \qquad [70]$$

When A^{-1} is factored out of the inverse matrix $\mathfrak{A}^{-1}$, Eq. 66, there is left

$$A\mathfrak{A}^{-1} = \mathfrak{A}^a = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} \qquad [71]$$

This matrix is called the *adjoint* of $\mathfrak{A}$. Since the adjoint contains only the cofactors of the elements of $\mathfrak{A}$, it may exist even when $\mathfrak{A}$ is singular, but only if the rank of $\mathfrak{A}$ is not less than $(n - 1)$ because all the cofactors would otherwise be zero.

According to Eq. 55, Ch. I, the determinant of $\mathfrak{A}^{-1}$ is A^{-1}; that is, the inverse matrix has a determinant which is the inverse of that of the given matrix. It is clear from Eq. 71 that the adjoint matrix $\mathfrak{A}^a$ of order n has a determinant which is A^n times that of $\mathfrak{A}^{-1}$. Hence the determinant of the adjoint is A^{n-1}; that is,

$$\begin{vmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{vmatrix} = A^{n-1} \qquad [72]$$

It follows that if the determinant of $\mathfrak{A}$ is not zero, the determinant of the

adjoint is not zero either, whereas if the determinant of $\mathfrak{A}$ is zero, that of $\mathfrak{A}^a$ is zero also.

Hence if $\mathfrak{A}$ has the rank n, then $\mathfrak{A}^a$ also has the rank n; but if $\mathfrak{A}$ has any rank less than $(n-1)$, the rank of $\mathfrak{A}^a$ is zero because all its elements are zero. When the rank of $\mathfrak{A}$ is $(n-1)$, at least one of the elements of $\mathfrak{A}^a$ is not zero and, therefore, its rank is at least 1. It can be shown to be no greater than 1, as is further discussed in Art. 8 of Ch. III.

The linear transformation 1 is for some purposes more effectively studied when written in the schematic form

$$
\begin{array}{c|ccccc}
 & x_1 & x_2 & x_3 & \cdots & x_n \\
\hline
y_1 & a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\
y_2 & a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\
y_3 & a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\
\cdot & & & & & \\
\cdot & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdot & & & & & \\
y_n & a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn}
\end{array}
\tag{73}
$$

which places the matrix more clearly in evidence. The Eqs. 1 are easily read out of this schematic arrangement by following along the rows and mentally dropping the x's down into their proper positions, supplying at the same time the necessary equal and plus signs. It now becomes more evident that these equations are closely related to another set, namely that which is obtained when the same manner of reading is carried out along the columns. Written out in the normal fashion, these equations are

$$
\left.\begin{array}{l}
a_{11}y_1 + a_{21}y_2 + \cdots + a_{n1}y_n = x_1 \\
a_{12}y_1 + a_{22}y_2 + \cdots + a_{n2}y_n = x_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{1n}y_1 + a_{2n}y_2 + \cdots + a_{nn}y_n = x_n
\end{array}\right\}
\tag{74}
$$

With reference to Eqs. 1, they are called the *transposed* set. Their matrix is

$$
\mathfrak{A}_t = \begin{bmatrix}
a_{11} & a_{21} & \cdots & a_{n1} \\
a_{12} & a_{22} & \cdots & a_{n2} \\
\cdots & \cdots & \cdots & \cdots \\
a_{1n} & a_{2n} & \cdots & a_{nn}
\end{bmatrix}
\tag{75}
$$

and is called the *transpose* (sometimes also the *conjugate*) of $\mathfrak{A}$. The transpose of $\mathfrak{A}$ is the matrix $\mathfrak{A}$ with its rows and columns interchanged.

In this book the transposed matrix is designated by the subscript t and the adjoint by the superscript a, as in Eqs. 75 and 71.

The transpose of a matrix evidently exists whether that matrix is

singular or not and also when the number of rows is different from the number of columns. In particular, the transpose of a row matrix is a column matrix, and vice versa; that is,

$$\lfloor x \rfloor_t = x] \tag{76}$$

and

$$x]_t = \lfloor x \rfloor \tag{77}$$

The transposed set of equations 74 should not be confused with the inverse set 52. For certain types of matrices the transpose and inverse are the same, but not in general.

In order for the inverse of a matrix to be equal to its transpose, the elements of that matrix must satisfy certain special conditions. It is easy to establish these conditions. If the elements of the matrix $\mathcal{A}$ are a_{rs}, those of the transpose $\mathcal{A}_t$ are a_{sr}, and the elements of the inverse matrix $\mathcal{A}^{-1}$ are given by Eq. 56. Hence the desired conditions are expressed by

$$a_{sr} = \frac{A_{sr}}{A} \tag{78}$$

The significance of this result may be made more evident by substituting it into the relations expressed by Eqs. 45 and 46 of Ch. I. These equations, it is recalled, hold generally for the elements and cofactors of any determinant. Substituting Eq. 78 into Eq. 45 of Ch. I gives

$$a_{1i}a_{1k} + a_{2i}a_{2k} + \cdots + a_{ni}a_{nk} = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \tag{79}$$

Substituting Eq. 78 into Eq. 46 of Ch. I yields the companion relation

$$a_{i1}a_{k1} + a_{i2}a_{k2} + \cdots + a_{in}a_{kn} = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \tag{80}$$

Equation 79 states that the sum of the squares of the elements of any column equals unity, but that the sum of the products of the elements of any column with the corresponding ones of any *other* column is equal to *zero*. Equation 80 expresses a similar relation with regard to the elements in the rows of the matrix. For example,

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2 = 1$$

or

$$a_{21}^2 + a_{22}^2 + \cdots + a_{n2}^2 = 1 \tag{81}$$

but

$$a_{11}a_{21} + a_{12}a_{22} + \cdots + a_{1n}a_{2n} = 0$$

or

$$a_{12}a_{13} + a_{22}a_{23} + \cdots + a_{n2}a_{n3} = 0 \qquad [82]$$

and so forth.

A matrix whose elements satisfy the relation 78, and hence the relations 79 and 80 also, is called an *orthogonal* matrix. The term "orthogonal" suggests the existence of a right-angle relationship. The geometrical interpretation which justifies the name for this particular type of matrix is given in the following chapter. A simple numerical illustration of an orthogonal matrix is

$$\mathcal{A} = \begin{bmatrix} 0.5 & 0.5 & 0.707 \\ -0.707 & 0.707 & 0 \\ -0.5 & -0.5 & 0.707 \end{bmatrix} \qquad [83]$$

The determinant of the matrix $\mathcal{A}$ is A. The determinant of its transpose is also A, but since the transpose of an orthogonal matrix is equal to its inverse, the determinant of $\mathcal{A}$ must at the same time be equal to A^{-1}. Hence the determinant of an orthogonal matrix must satisfy the equation

$$A = \frac{1}{A} \qquad [84]$$

whence

$$A^2 = 1 \qquad [85]$$

and

$$A = \pm 1 \qquad [86]$$

The conditions under which the algebraic sign is either plus or minus are discussed subsequently. For the matrix 83, the determinant is $+1$, as the reader may readily verify.

The transpose of the inverse of a matrix plays a sufficiently important role in the subjects utilizing matrix algebra to justify for it a special name and designation. It is called the *reciprocal* matrix, and it is designated by an asterisk affixed to the corresponding script letter. Thus the reciprocal of $\mathcal{A}$ is given by

$$\mathcal{A}^* = \mathcal{A}_t^{-1} \qquad [87]$$

The elements of the reciprocal matrix are similarly designated by an asterisk placed on the elements a_{rs} of $\mathcal{A}$. Reference to Eq. 66 shows that the elements of $\mathcal{A}^*$ are given by

$$a^*_{rs} = \frac{A_{rs}}{A} \qquad [88]$$

The same result is obtained by calculation of the inverse of the transpose, so transposition and inversion are commutative.

The following summary may be helpful at this point:

For a given matrix $\mathcal{A}$ with elements a_{rs}, determinant A, and cofactors A_{rs}:

The inverse of $\mathcal{A}$ is $\mathcal{A}^{-1}$ with elements $\dfrac{A_{sr}}{A}$ and determinant A^{-1}.

The adjoint of $\mathcal{A}$ is $\mathcal{A}^{a}$ with elements A_{sr} and determinant A^{n-1}.

The transpose of $\mathcal{A}$ is $\mathcal{A}_t$ with elements a_{sr} and determinant A.

The reciprocal of $\mathcal{A}$ is $\mathcal{A}^*$ with elements $a^*_{rs} = \dfrac{A_{rs}}{A}$ and determinant A^{-1}.

If $\mathcal{A}$ is orthogonal, then $\mathcal{A}_t = \mathcal{A}^{-1}$, $a_{rs} = \dfrac{A_{rs}}{A}$, and determinant $A = \pm 1$. Also $\mathcal{A}^* = \mathcal{A}$; that is, an orthogonal matrix is its own reciprocal.

In the manipulation of matrix equations it is sometimes useful to note that

$$(\mathcal{A} \times \mathcal{B})_t = \mathcal{B}_t \times \mathcal{A}_t \tag{89}$$

that is, the transpose of a product is equal to the *reversed* product of the individually transposed matrices. This conclusion follows from the fact that the product $(\mathcal{A} \times \mathcal{B})_t$ is formed by multiplying the rows of $\mathcal{A}$ by the columns of $\mathcal{B}$ and then interchanging the rows and columns in the resulting matrix, which is the same as though the elements had been formed by multiplying the columns of $\mathcal{B}$ (these are the rows of $\mathcal{B}_t$) by the rows of $\mathcal{A}$ (these are the columns of $\mathcal{A}_t$). For example, the (2,1)-element in the resultant matrix $(\mathcal{A} \times \mathcal{B})_t$ is formed from the first row in $\mathcal{A}$ and the second column in $\mathcal{B}$; and the (2,1)-element in the resultant matrix $\mathcal{B}_t \times \mathcal{A}_t$ is formed from the second row in $\mathcal{B}_t$ (second column in $\mathcal{B}$) and the first column in $\mathcal{A}_t$ (first row in $\mathcal{A}$), which are identical combinations.

This relationship may readily be extended to multiple products. If

$$\mathcal{A} \times \mathcal{B} = \mathcal{G} \tag{90}$$

then

$$(\mathcal{A} \times \mathcal{B})_t = \mathcal{G}_t = \mathcal{B}_t \times \mathcal{A}_t \tag{91}$$

Next if

$$\mathcal{A} \times \mathcal{B} \times \mathcal{C} = \mathcal{G} \times \mathcal{C} \tag{92}$$

then

$$(\mathcal{A} \times \mathcal{B} \times \mathcal{C})_t = (\mathcal{G} \times \mathcal{C})_t = \mathcal{C}_t \times \mathcal{G}_t = \mathcal{C}_t \times \mathcal{B}_t \times \mathcal{A}_t \tag{93}$$

In general,

$$(\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \cdots \times \mathcal{P})_t = \mathcal{P}_t \times \cdots \times \mathcal{C}_t \times \mathcal{B}_t \times \mathcal{A}_t \qquad [94]$$

The transpose of a multiple product is equal to the reversed product of the individually transposed matrices.

A similar relationship also holds for the *inverse* of a matrix product. If

$$\mathcal{A} \times \mathcal{B} = \mathcal{C} \qquad [95]$$

then from the fact that

$$\mathcal{C} \times \mathcal{C}^{-1} = \mathcal{U} \qquad \text{(unit matrix)} \qquad [96]$$

and

$$\mathcal{A} \times \mathcal{B} \times \mathcal{B}^{-1} \times \mathcal{A}^{-1} = \mathcal{U} \qquad [97]$$

it is seen that

$$\mathcal{C}^{-1} = \mathcal{B}^{-1}\mathcal{A}^{-1} \qquad [98]$$

Hence

$$(\mathcal{A} \times \mathcal{B})^{-1} = \mathcal{B}^{-1} \times \mathcal{A}^{-1} \qquad [99]$$

The inverse of a product is equal to the reversed product of the individual inverse matrices.

The extension to multiple products follows as before,

$$(\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \cdots \times \mathcal{P})^{-1} = \mathcal{P}^{-1} \times \cdots \times \mathcal{C}^{-1} \times \mathcal{B}^{-1} \times \mathcal{A}^{-1} \qquad [100]$$

It is clear that all the matrices must be nonsingular in order for this relation to apply, but this fact does not mean that a multiple product involving singular matrices cannot in some cases have an inverse. For example, the component matrices in the following product are evidently singular because they are not square:

$$[a_{24}] \times [b_{42}] = [c_{22}] \qquad [101]$$

But the resultant matrix is square and may be nonsingular, in which case it possesses an inverse. The relation 99 is not applicable to the product in Eq. 101.

Similar reasoning shows that an analogous relationship holds also for the adjoint of a product of two matrices

$$(\mathcal{A} \times \mathcal{B})^a = \mathcal{B}^a \times \mathcal{A}^a \qquad [102]$$

as well as for the adjoint of a multiple product

$$(\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \cdots \times \mathcal{P})^a = \mathcal{P}^a \times \cdots \times \mathcal{C}^a \times \mathcal{B}^a \times \mathcal{A}^a \qquad [103]$$

According to the definition (Eq. 87) of the reciprocal of a matrix, it follows from Eqs. 89 and 99 that

$$(\mathcal{A} \times \mathcal{B})^* = \mathcal{A}^* \times \mathcal{B}^* \qquad [104]$$

The reciprocal of the product of two matrices is equal to the product (*not* reversed) of the individual reciprocals. The extension of this rule to multiple products evidently reads

$$(\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \cdots \times \mathcal{P})^* = \mathcal{A}^* \times \mathcal{B}^* \times \mathcal{C}^* \times \cdots \times \mathcal{P}^* \qquad [105]$$

The relations expressed by Eqs. 89 and 99 also show that if $\mathcal{P}$ and $\mathcal{Q}$ are orthogonal matrices, so that

$$\mathcal{P}_t = \mathcal{P}^{-1} \qquad \text{and} \qquad \mathcal{Q}_t = \mathcal{Q}^{-1} \qquad [106]$$

then

$$(\mathcal{P} \times \mathcal{Q})_t = (\mathcal{P} \times \mathcal{Q})^{-1} \qquad [107]$$

Hence *the product of two orthogonal matrices is again orthogonal.*

7. Submatrices or the partitioning of matrices

In carrying out a matrix product, subdividing or partitioning the matrices into smaller components is sometimes convenient. In the following example

$$\mathcal{A} \times \mathcal{B} = \begin{array}{c} \longleftarrow r \longrightarrow \quad \\ \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{array}\right] \end{array} \times \begin{array}{cc} \left[\begin{array}{c|cc} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ \hline b_{31} & b_{32} & b_{33} \end{array}\right] & \begin{array}{c} \updownarrow r \\ \\ \end{array} \end{array} \qquad [108]$$

the so-called *submatrices* in $\mathcal{A}$ are

$$\alpha_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \alpha_{12} = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}$$
$$\alpha_{21} = \begin{bmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \qquad \alpha_{22} = \begin{bmatrix} a_{33} \\ a_{43} \end{bmatrix} \qquad [109]$$

and in $\mathcal{B}$ they are

$$\beta_{11} = \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \qquad \beta_{12} = \begin{bmatrix} b_{12} & b_{13} \\ b_{22} & b_{23} \end{bmatrix}$$
$$\beta_{21} = \begin{bmatrix} b_{31} \end{bmatrix} \qquad \beta_{22} = \begin{bmatrix} b_{32} & b_{33} \end{bmatrix} \qquad [110]$$

The matrix product, Eq. 108, may be written

$$\mathcal{A} \times \mathcal{B} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \times \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \qquad [111]$$

and evaluated as though the submatrices were ordinary elements except that the order in which they enter into the products formed by this

evaluation process must be observed carefully. Such products of submatrices must, of course, subsequently be carried out by the same rule of matrix multiplication.

When the original matrices are partitioned as is done in Eq. 108, the division of the columns of $\mathfrak{A}$ into subgroups must be identical with the division of the rows of $\mathfrak{B}$ into subgroups. Thus the subgroup of the first r columns of $\mathfrak{A}$ must be matched by a subgroup of the first r rows of $\mathfrak{B}$. This matching assures that all the submatrices appearing in the products formed during the subsequent evaluation of Eq. 111 will be conformable.

If the total number of columns in $\mathfrak{A}$ are subdivided into more than two subgroups, the rows of $\mathfrak{B}$ must be similarly subdivided. The rows of $\mathfrak{A}$ and the columns of $\mathfrak{B}$, on the other hand, may be subdivided in any desired manner without affecting the conformability of submatrices later appearing in the evaluation process. In Eq. 108 the columns of the matrix $\mathfrak{B}$ may, for example, not be subdivided at all or they may be divided into three subgroups instead of two.

It is an instructive exercise to show, by means of the relations 37, that the total operations of multiplication and of addition involved in the complete evaluation of a matrix product are independent of whether or how the component matrices are subdivided. This fact alone might be regarded as an indication that the present discussion is, for practical purposes, wholly irrelevant except that subdivision may be convenient in certain special cases. A somewhat different application of the principle of subdividing matrices is, however, very useful practically in the solution of a set of linear equations or in the inversion of a matrix.

The subdivided matrix equation for a linear transformation is indicated by

$$\left[\begin{array}{cc:ccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \hdashline a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array}\right] \times \left[\begin{array}{c} x_1 \\ x_2 \\ \hdashline x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{array}\right] = \left[\begin{array}{c} y_1 \\ y_2 \\ \hdashline y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{array}\right] \qquad [112]$$

Denoting the submatrices by

$$\alpha_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \alpha_{12} = \begin{bmatrix} a_{13} & \cdots & a_{1n} \\ a_{23} & \cdots & a_{2n} \end{bmatrix}$$

$$\alpha_{21} = \begin{bmatrix} a_{31} & a_{32} \\ \cdots & \cdots \\ a_{n1} & a_{n2} \end{bmatrix} \qquad \alpha_{22} = \begin{bmatrix} a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots \\ a_{n3} & \cdots & a_{nn} \end{bmatrix} \qquad [113]$$

and

$$\xi_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \xi_2 = \begin{bmatrix} x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$
$$\eta_1 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \eta_2 = \begin{bmatrix} y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \tag{114}$$

one may write the transformation, Eq. 112, as

$$\left.\begin{aligned} \alpha_{11}\xi_1 + \alpha_{12}\xi_2 &= \eta_1 \\ \alpha_{21}\xi_1 + \alpha_{22}\xi_2 &= \eta_2 \end{aligned}\right\} \tag{115}$$

A possible method of solution is the following. The second equation in the set 115 may be premultiplied by α_{22}^{-1} and solved for ξ_2, giving

$$\xi_2 = \alpha_{22}^{-1}\eta_2 - \alpha_{22}^{-1}\alpha_{21}\xi_1 \tag{116}$$

Substituting this into the first equation of set 115 gives

$$(\alpha_{11} - \alpha_{12}\alpha_{22}^{-1}\alpha_{21})\xi_1 = \eta_1 - \alpha_{12}\alpha_{22}^{-1}\eta_2 \tag{117}$$

from which

$$\xi_1 = (\alpha_{11} - \alpha_{12}\alpha_{22}^{-1}\alpha_{21})^{-1}(\eta_1 - \alpha_{12}\alpha_{22}^{-1}\eta_2) \tag{118}$$

In a similar fashion one obtains the solution for ξ_2,

$$\xi_2 = (\alpha_{22} - \alpha_{21}\alpha_{11}^{-1}\alpha_{12})^{-1}(\eta_2 - \alpha_{21}\alpha_{11}^{-1}\eta_1) \tag{119}$$

As is to be expected, this is simply Eq. 118 with the subscripts 1 and 2 interchanged.

Writing for brevity

$$\begin{aligned} \theta_1 &= (\alpha_{11} - \alpha_{12}\alpha_{22}^{-1}\alpha_{21}) \\ \theta_2 &= (\alpha_{22} - \alpha_{21}\alpha_{11}^{-1}\alpha_{12}) \end{aligned} \tag{120}$$

the inverse of the matrix $\mathfrak{A}$ in the transformation 112 is indicated by

$$\mathfrak{A}^{-1} = \left[\begin{array}{c:c} \theta_1^{-1} & -\theta_1^{-1}\alpha_{12}\alpha_{22}^{-1} \\ \hdashline -\theta_2^{-1}\alpha_{21}\alpha_{11}^{-1} & \theta_2^{-1} \end{array}\right] \tag{121}$$

The inversion of a matrix of given order is thus reduced to the inversion of matrices of lower order. These are the submatrices α_{11} and α_{22}, and the resultant matrices θ_1 and θ_2 given by Eqs. 120. For α_{11} and α_{22} to be

nonsingular, they must have the same number of rows as columns. The matrix $\mathcal{A}$ must, therefore, be subdivided to meet this condition, as is done in Eq. 112. Even then it may happen that α_{11} or α_{22} is singular, but this situation can always be remedied by a rearrangement in the order of the original equations.

The submatrices α_{12} and α_{21} do not need to be inverted. They may have any number of rows irrespective of the number of columns. The matrices θ_1 and θ_2 are evidently square and of the same order as α_{11} and α_{22} respectively. The process is, therefore, always applicable provided, of course, that $\mathcal{A}$ is nonsingular.

When the matrix $\mathcal{A}$ has many rows and columns, an extension of the method may be developed which allows for further subdivision, or the present method may again be applied to some of the submatrices obtained from a preliminary subdivision.

A numerical example may serve to illustrate the advantage of this method over the one suggested by Cramer's rule. Let the matrix be

$$\mathcal{A} = \left[\begin{array}{cc:cc} 2 & 3 & -4 & 1 \\ 1 & -2 & 2 & 4 \\ \hdashline 3 & 2 & -2 & 3 \\ -2 & 1 & 3 & -1 \end{array}\right] \tag{122}$$

For the indicated subdivision,

$$\begin{aligned} \alpha_{11} &= \begin{bmatrix} 2 & 3 \\ 1 & -2 \end{bmatrix} & \alpha_{12} &= \begin{bmatrix} -4 & 1 \\ 2 & 4 \end{bmatrix} \\ \alpha_{21} &= \begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix} & \alpha_{22} &= \begin{bmatrix} -2 & 3 \\ 3 & -1 \end{bmatrix} \end{aligned} \tag{123}$$

In this simple case the submatrices α_{11} and α_{22} may be inverted by inspection, giving

$$\alpha_{11}^{-1} = \frac{1}{7}\begin{bmatrix} 2 & 3 \\ 1 & -2 \end{bmatrix} \qquad \alpha_{22}^{-1} = \frac{1}{7}\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{124}$$

Next in order is the determination of

$$\alpha_{12}\alpha_{22}^{-1} = \frac{1}{7}\begin{bmatrix} -4 & 1 \\ 2 & 4 \end{bmatrix} \times \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} -1 & -10 \\ 14 & 14 \end{bmatrix} \tag{125}$$

and

$$\alpha_{21}\alpha_{11}^{-1} = \frac{1}{7}\begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 3 \\ 1 & -2 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 8 & 5 \\ -3 & -8 \end{bmatrix} \tag{126}$$

Then

$$\alpha_{12}\alpha_{22}^{-1}\alpha_{21} = \frac{1}{7}\begin{bmatrix} -1 & -10 \\ 14 & 14 \end{bmatrix} \times \begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 17 & -12 \\ 14 & 42 \end{bmatrix} \qquad [127]$$

$$\alpha_{21}\alpha_{11}^{-1}\alpha_{12} = \frac{1}{7}\begin{bmatrix} 8 & 5 \\ -3 & -8 \end{bmatrix} \times \begin{bmatrix} -4 & 1 \\ 2 & 4 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} -22 & 28 \\ -4 & -35 \end{bmatrix} \qquad [128]$$

whence

$$\theta_1 = \begin{bmatrix} 2 & 3 \\ 1 & -2 \end{bmatrix} - \frac{1}{7}\begin{bmatrix} 17 & -12 \\ 14 & 42 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} -3 & 33 \\ -7 & -56 \end{bmatrix} \qquad [129]$$

$$\theta_1^{-1} = 7\begin{bmatrix} -3 & 33 \\ -7 & -56 \end{bmatrix}^{-1} = \frac{1}{57}\begin{bmatrix} -56 & -33 \\ 7 & -3 \end{bmatrix} \qquad [130]$$

and

$$\theta_2 = \begin{bmatrix} -2 & 3 \\ 3 & -1 \end{bmatrix} - \frac{1}{7}\begin{bmatrix} -22 & 28 \\ -4 & -35 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 8 & -7 \\ 25 & 28 \end{bmatrix} \qquad [131]$$

$$\theta_2^{-1} = 7\begin{bmatrix} 8 & -7 \\ 25 & 28 \end{bmatrix}^{-1} = \frac{1}{57}\begin{bmatrix} 28 & 7 \\ -25 & 8 \end{bmatrix} \qquad [132]$$

Finally,

$$-\theta_1^{-1}\alpha_{12}\alpha_{22}^{-1} = \frac{1}{57}\begin{bmatrix} 56 & 33 \\ -7 & 3 \end{bmatrix} \times \frac{1}{7}\begin{bmatrix} -1 & -10 \\ 14 & 14 \end{bmatrix} = \frac{1}{57}\begin{bmatrix} 58 & -14 \\ 7 & 16 \end{bmatrix} \qquad [133]$$

and

$$-\theta_2^{-1}\alpha_{21}\alpha_{11}^{-1} = \frac{1}{57}\begin{bmatrix} 28 & 7 \\ -25 & 8 \end{bmatrix} \times \frac{1}{7}\begin{bmatrix} -8 & -5 \\ 3 & 8 \end{bmatrix} = \frac{1}{57}\begin{bmatrix} -29 & -12 \\ 32 & 27 \end{bmatrix} \qquad [134]$$

When Eqs. 130, 132, 133, and 134 are put together according to the form indicated in Eq. 121, the desired inverse is found to be

$$\mathfrak{A}^{-1} = \frac{1}{57}\begin{bmatrix} -56 & -33 & 58 & -14 \\ 7 & -3 & 7 & 16 \\ -29 & -12 & 28 & 7 \\ 32 & 27 & -25 & 8 \end{bmatrix} \qquad [135]$$

The student should obtain this same result by means of Cramer's rule directly in order to appreciate how considerable a saving in labor is afforded by the use of submatrices.

The matrix of Eqs. 112 may be manipulated in a variety of additional ways which yield slight modifications in the process of obtaining the same end result, but there is little point in further discussing this item here since the fundamental principle remains the same.

When, after subdivision, the given matrix has the form

$$\mathcal{A} = \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \tag{136}$$

Eqs. 120 and 121 show that the inverse is given by

$$\mathcal{A}^{-1} = \begin{bmatrix} \alpha_{11}^{-1} & 0 \\ 0 & \alpha_{22}^{-1} \end{bmatrix} \tag{137}$$

On the other hand, if

$$\mathcal{A} = \begin{bmatrix} 0 & \alpha_{12} \\ \alpha_{21} & 0 \end{bmatrix} \tag{138}$$

the inversion of Eq. 115 shows that

$$\mathcal{A}^{-1} = \begin{bmatrix} 0 & \alpha_{21}^{-1} \\ \alpha_{12}^{-1} & 0 \end{bmatrix} \tag{139}$$

These simple examples are readily generalized, with the result that if

$$\mathcal{A} = \begin{bmatrix} \alpha_{11} & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_{nn} \end{bmatrix} \tag{140}$$

then

$$\mathcal{A}^{-1} = \begin{bmatrix} \alpha_{11}^{-1} & 0 & \cdots & 0 \\ 0 & \alpha_{22}^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_{nn}^{-1} \end{bmatrix} \tag{141}$$

and if

$$\mathcal{A} = \begin{bmatrix} 0 & \cdots & 0 & \alpha_{1n} \\ 0 & \cdots & \alpha_{2,n-1} & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \cdots & 0 & 0 \end{bmatrix} \tag{142}$$

then the inverse is given by

$$\mathcal{A}^{-1} = \begin{bmatrix} 0 & \cdots & 0 & \alpha_{n1}^{-1} \\ 0 & \cdots & \alpha^{-1}{}_{n-1,2} & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{1n}^{-1} & \cdots & 0 & 0 \end{bmatrix} \tag{143}$$

8. The linear transformation of matrices

In Art. 1 it is shown that if the variables $y_1 \cdots y_n$ in the transformation 1 are subjected to the further transformation 14, the matrix which com-

bines the two transformations into one is given by the matrix product expressed by Eq. 20. In this process the matrix $\mathcal{A}$ is said to be *linearly transformed* into the matrix $\mathcal{C}$.

The matrix $\mathcal{B}$, by means of which the transformation of $\mathcal{A}$ is effected, may sometimes advantageously be thought of as the resultant of a multiple product of very simple component matrices; thus

$$\mathcal{B} = \mathcal{B}_n \times \cdots \times \mathcal{B}_3 \times \mathcal{B}_2 \times \mathcal{B}_1 \qquad [144]$$

The transformation of $\mathcal{A}$ is then considered to be accomplished through the succession of a number of simpler transformations, the first of which is effected by $\mathcal{B}_1$, the second by $\mathcal{B}_2$, and so forth.

A matrix may be transformed by postmultiplication as well as by premultiplication, in which case the matrix effecting the transformation is decomposed into components according to the order indicated by

$$\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3 \times \cdots \times \mathcal{B}_n \qquad [145]$$

Here $\mathcal{B}_1$ accomplishes the first simple transformation, $\mathcal{B}_2$ the second, and so forth.

It is of interest to determine the simplest fundamental forms into which an arbitrary nonsingular matrix $\mathcal{B}$ may be decomposed, and to interpret the individual transformations which they effect upon the form of a matrix $\mathcal{A}$. They are called *elementary transformations*, and the matrices which produce them are the *elementary transformation matrices*. There are three types of elementary transformations, and, correspondingly, there are three fundamental types of so-called elementary transformation matrices.

The first type of elementary transformation amounts to an interchange of any two rows or columns of the matrix $\mathcal{A}$. The second type is the addition of the elements of a row or column to the corresponding elements of another row or column; and the third type is the multiplication of any row or column of $\mathcal{A}$ by an arbitrary nonzero factor. Each type of transformation may be effected through multiplying $\mathcal{A}$ by a corresponding type of transformation matrix. If the desired transformation is intended to affect the *rows*, $\mathcal{A}$ is *pre*multiplied by the transformation matrix; if the *columns* are to be affected, $\mathcal{A}$ is *post*multiplied by the transformation matrix.

Each type of elementary transformation matrix is formed from the unit matrix $\mathcal{U}$, Eq. 47, by performing upon it the *same* elementary transformation which the desired transformation matrix is intended to effect in the matrix $\mathcal{A}$ by means of pre- or postmultiplication. Thus $\mathcal{U}$ with its pth and qth rows (respectively columns) interchanged yields a transformation matrix (type 1) which, by means of pre- (respectively post-) multiplication, interchanges the pth and qth rows (respectively

columns) of $\mathcal{A}$. The matrix $\mathcal{U}$ with its qth row (respectively column) added to its pth row (respectively column) yields the transformation matrix (type 2) which, by means of pre- (respectively post-) multiplication, effects the addition of the elements of the qth row (respectively column) of $\mathcal{A}$ to the corresponding elements of its pth row (respectively column). Finally, the matrix $\mathcal{U}$ with its pth row (respectively column) multiplied by a factor k yields a transformation matrix (type 3) which, by means of pre- (respectively post-) multiplication, multiplies the elements of the pth row (respectively column) of $\mathcal{A}$ by k.

Observe that each type of elementary transformation matrix has two forms according to whether it is intended to effect a transformation of the rows or columns of $\mathcal{A}$ (by pre- or postmultiplication, respectively). The one form in each case is evidently the transpose of the other, and in types 1 and 3 these two are readily seen to be identical, whereas in type 2 the transposed matrix merely interchanges the distinction between the designations p and q in the description of the preceding paragraph.

The elementary transformation matrices are in this text denoted by the script letter $\mathcal{T}$ with subscripts intended to indicate the type. Thus $\mathcal{T}_{p\sim q}$ is used to designate type 1; $\mathcal{T}_{\underline{p+q}}$ designates type 2 for the transformation of rows, whereas $\mathcal{T}_{p+q|}$ is the designation of this type for the transformation of columns; and $\mathcal{T}_{p\times k}$ denotes the transformation matrix of type 3.

These are illustrated for the case of fourth-order matrices by the following examples:

$$\mathcal{T}_{1\sim 3} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = (\mathcal{T}_{1\sim 3})_t \quad \text{(type 1)} \tag{146}$$

$$\mathcal{T}_{\underline{2+3}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = (\mathcal{T}_{2+3|})_t \quad \text{(type 2 for rows)} \tag{147}$$

$$\mathcal{T}_{2+3|} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = (\mathcal{T}_{\underline{2+3}})_t \quad \text{(type 2 for columns)} \tag{148}$$

$$\mathcal{T}_{3\times k} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & k & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = (\mathcal{T}_{3\times k})_t \quad \text{(type 3)} \tag{149}$$

The transformation of a fourth-order matrix $\mathcal{A}$ by means of these is illustrated in the numerical examples given below:

$$\mathcal{T}_{1\sim 3}\times\mathcal{A}=\begin{bmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{bmatrix}\times\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}=\begin{bmatrix}4&2&6&9\\1&5&8&7\\2&4&3&6\\10&5&3&1\end{bmatrix} \quad [150]$$

$$\mathcal{A}\times\mathcal{T}_{1\sim 3}=\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}\times\begin{bmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{bmatrix}=\begin{bmatrix}3&4&2&6\\8&5&1&7\\6&2&4&9\\3&5&10&1\end{bmatrix} \quad [151]$$

$$\mathcal{T}_{\underline{2+3}}\times\mathcal{A}=\begin{bmatrix}1&0&0&0\\0&1&1&0\\0&0&1&0\\0&0&0&1\end{bmatrix}\times\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}=\begin{bmatrix}2&4&3&6\\5&7&14&16\\4&2&6&9\\10&5&3&1\end{bmatrix} \quad [152]$$

$$\mathcal{A}\times\mathcal{T}_{2+3|}=\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}\times\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&1&1&0\\0&0&0&1\end{bmatrix}=\begin{bmatrix}2&7&3&6\\1&13&8&7\\4&8&6&9\\10&8&3&1\end{bmatrix} \quad [153]$$

$$\mathcal{T}_{3\times(5)}\times\mathcal{A}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&5&0\\0&0&0&1\end{bmatrix}\times\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}=\begin{bmatrix}2&4&3&6\\1&5&8&7\\20&10&30&45\\10&5&3&1\end{bmatrix} \quad [154]$$

$$\mathcal{A}\times\mathcal{T}_{3\times(5)}=\begin{bmatrix}2&4&3&6\\1&5&8&7\\4&2&6&9\\10&5&3&1\end{bmatrix}\times\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&5&0\\0&0&0&1\end{bmatrix}=\begin{bmatrix}2&4&15&6\\1&5&40&7\\4&2&30&9\\10&5&15&1\end{bmatrix} \quad [155]$$

The transformation matrix of type 2 may be generalized from the form $\mathcal{T}_{p+q}$ to $\mathcal{T}_{p\pm q}$ (with reference to either rows or columns), which allows the qth row or column to be subtracted from as well as added to the pth row or column. It is useful to note that for this type

$$\mathcal{T}_{\underline{p\pm q}} = \mathcal{T}_{q\pm p|} \quad [156]$$

and since

$$(\mathcal{T}_{\underline{p\pm q}})_t = \mathcal{T}_{p\pm q|} \quad [157]$$

it follows that

$$(\mathcal{T}_{\underline{p\pm q}})_t = \mathcal{T}_{\underline{q\pm p}} \quad [158]$$

and

$$(\mathcal{T}_{p\pm q|})_t = \mathcal{T}_{q\pm p|} \quad [159]$$

For the matrices of types 1 and 3 it is clear that

$$(\mathcal{T}_{p\sim q})_t = \mathcal{T}_{p\sim q} \tag{160}$$

and

$$(\mathcal{T}_{p\times k})_t = \mathcal{T}_{p\times k} \tag{161}$$

It is also useful to note the form of the inverse of each type of transformation matrix, thus

$$\mathcal{T}_{p\sim q}{}^{-1} = \mathcal{T}_{p\sim q} \tag{162}$$

$$\mathcal{T}_{p\pm q}{}^{-1} = \mathcal{T}_{p\mp q} \qquad \text{(row or column)} \tag{163}$$

and

$$\mathcal{T}_{p\times k}{}^{-1} = \mathcal{T}_{p\times k^{-1}} \tag{164}$$

According to Eqs. 160 and 162, the transpose of the matrix of type 1 equals its inverse. This type of transformation matrix is therefore orthogonal.

In the transformation of matrices it is useful to have a type of transformation matrix which effects the addition (or subtraction) of the k-multiplied qth row or column to (or from) the pth row or column in a single operation. Such a matrix is evidently the following combination of types 2 and 3:

$$\mathcal{T}_{\underline{p\pm kq}} = \mathcal{T}_{q\times k^{-1}} \times \mathcal{T}_{\underline{p\pm q}} \times \mathcal{T}_{q\times k} \tag{165}$$

or

$$\mathcal{T}_{p\pm kq|} = \mathcal{T}_{q\times k} \times \mathcal{T}_{p\pm q|} \times \mathcal{T}_{q\times k^{-1}} \tag{166}$$

according to whether the rows are to be affected by premultiplication or the columns by postmultiplication, respectively. The matrix $\mathcal{T}_{\underline{p\pm kq}}$ or $\mathcal{T}_{p\pm kq|}$ is formed from the unit matrix $\mathcal{U}$ by adding (or subtracting) its k-multiplied qth row or column to (or from) the pth row or column, respectively. It may be referred to as type 23. It has the properties

$$(\mathcal{T}_{\underline{p\pm kq}})_t = \mathcal{T}_{p\pm kq|} \tag{167}$$

$$\mathcal{T}_{p\pm kq}{}^{-1} = \mathcal{T}_{p\mp kq} \qquad \text{(row or column)} \tag{168}$$

A numerical illustration of the use of this matrix is

$$\mathcal{T}_{\underline{2+(4)3}} \times \mathcal{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 4 & 3 & 6 \\ 1 & 5 & 8 & 7 \\ 4 & 2 & 6 & 9 \\ 10 & 5 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 3 & 6 \\ 17 & 13 & 32 & 43 \\ 4 & 2 & 6 & 9 \\ 10 & 5 & 3 & 1 \end{bmatrix} \tag{169}$$

$$\mathcal{A} \times \mathcal{T}_{2+(4)3|} = \begin{bmatrix} 2 & 4 & 3 & 6 \\ 1 & 5 & 8 & 7 \\ 4 & 2 & 6 & 9 \\ 10 & 5 & 3 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 16 & 3 & 6 \\ 1 & 37 & 8 & 7 \\ 4 & 26 & 6 & 9 \\ 10 & 17 & 3 & 1 \end{bmatrix} \tag{170}$$

According to the theory of determinants, it is clear that the transformation matrices have determinants with the following values:

$$\text{(type 1)} \quad T_{p\sim q} = -1 \qquad [171]$$

$$\text{(type 2)} \quad T_{p\pm q} = 1 \quad \text{(row or column)} \qquad [172]$$

$$\text{(type 3)} \quad T_{p\times k} = k \qquad [173]$$

$$\text{(type 23)} \quad T_{p\pm kq} = 1 \quad \text{(row or column)} \qquad [174]$$

Hence transformation of the matrix $\mathcal{A}$ by means of types 2 or 23 leaves the value of the determinant A unchanged; transformation by means of type 1 merely changes the algebraic sign of A; and transformation by means of type 3 has the effect of multiplying the determinant A by the factor k.

It should be observed that an elementary transformation matrix $\mathcal{T}$ is always square (has the same number of rows as columns), but $\mathcal{A}$ need not be. However, the order of $\mathcal{T}$ must be so chosen that the matrices are conformable. If $\mathcal{A}$ is not square, the correct order of $\mathcal{T}$ is different in the cases of pre- and postmultiplication.

The three types of elementary transformation with the matrices $\mathcal{T}_{p\sim q}$, $\mathcal{T}_{p\pm q}$, and $\mathcal{T}_{p\times k}$, if repeatedly applied to an arbitrary matrix $\mathcal{A}$, are capable of yielding the same end result as is expressed by the relation

$$\mathcal{P} \times \mathcal{A} \times \mathcal{Q} = \mathcal{C} \qquad [175]$$

in which $\mathcal{P}$ and $\mathcal{Q}$ are any nonsingular matrices. This is the same as saying that any nonsingular matrix such as $\mathcal{P}$ or $\mathcal{Q}$ can be represented as a product of suitably chosen elementary matrices of types 1, 2, and 3 alone.

9. Equivalence of matrices

A matrix $\mathcal{C}$ which may be obtained from a matrix $\mathcal{A}$ by means of a finite number of elementary transformations is said to be *equivalent* to $\mathcal{A}$. The equivalence of matrices is evidently a mutual relationship, since the elementary transformations are nonsingular and hence reversible. Thus if $\mathcal{C}$ can be obtained from $\mathcal{A}$ by a succession of elementary transformations, it follows that $\mathcal{A}$ can be regained by means of elementary transformations. The state of equivalence of two matrices $\mathcal{A}$ and $\mathcal{C}$ may evidently be expressed by Eq. 175, in which $\mathcal{P}$ and $\mathcal{Q}$ are arbitrary nonsingular matrices.

It is important for later applications to recognize that equivalent matrices have the same rank. The truth of this statement may be seen to follow from the fact that the rank of a matrix depends upon properties of the corresponding determinant (the vanishing or nonvanishing of this

determinant and its minors), which, according to the theory of determinants, are not altered by the elementary transformations.

10. Transformation of a square matrix to the diagonal form

A square matrix possesses equivalent diagonal forms, any of which may be obtained through a linear transformation having the form of Eq. 175. Such a transformation may be written

$$\mathcal{P} \times \mathcal{A} \times \mathcal{Q} = \mathcal{D} \qquad [176]$$

$\mathcal{D}$ being an equivalent diagonal form of $\mathcal{A}$.

The matrices $\mathcal{P}$ and $\mathcal{Q}$ which accomplish the transformation can be formed from components of the combination type 23 alone (see Art. 8), in which case $\mathcal{A}$ and $\mathcal{D}$ have the same determinant.

The detailed process is best shown by means of a numerical example. The one given in Art. 2, Ch. I, illustrating the numerical evaluation of a determinant, involves precisely the elementary transformations required here. The given matrix $\mathcal{A}$ may be assumed to have the determinant A expressed by Eq. 4, Ch. I. The matrix $\mathcal{P}$ consists of components of the type $\mathcal{T}_{p \pm kq}$, their formation being described in Art. 2, Ch. I, by steps 1, 2, and 3. Since premultiplication is required by these transformations, the order of the components referring respectively to these steps reads from right to left, thus

$$\mathcal{P} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{8}{10} & 1 \end{bmatrix}}_{\text{Third step}} \times \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}}_{\text{Second step}} \times \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{First step}} = \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix} \qquad [177]$$

Similarly, the matrix $\mathcal{Q}$ is given by the product of components of the type $\mathcal{T}_{p \pm kq'}$, their formation being described by steps 4, 5, and 6. Here postmultiplication is involved. Hence the order of the component matrices corresponding respectively to these steps reads from left to right. The desired matrix $\mathcal{Q}$ is, therefore, given by

$$\mathcal{Q} = \underbrace{\begin{bmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Fourth step}} \times \underbrace{\begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Fifth step}} \times \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{2}{10} \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Sixth step}} = \begin{bmatrix} 1 & -3 & -\frac{7}{5} \\ 0 & 1 & -\frac{1}{5} \\ 0 & 0 & 1 \end{bmatrix} \qquad [178]$$

The transformation expressed by Eq. 176 reads

$$\begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 2 \\ 4 & 2 & 6 \\ 3 & 1 & 7 \end{bmatrix} \times \begin{bmatrix} 1 & -3 & -\frac{7}{5} \\ 0 & 1 & -\frac{1}{5} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} \qquad [179]$$

The determinant A is equal to the product of the diagonal elements.

It is obvious that, by means of elementary transformations of the type $\mathcal{T}_{p\times k}$, this resultant diagonal matrix may be made to go over into the unit matrix of like order.

It is also possible to transform a matrix $\mathcal{A}$ to the diagonal form by premultiplication or postmultiplication alone. For example, after the matrix is reduced to the triangular form in the manner described above, one has

$$\mathcal{P} \times \mathcal{A} = \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 2 \\ 4 & 2 & 6 \\ 3 & 1 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 2 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} \qquad [180]$$

The elements above the principal diagonal in this triangular matrix may now be reduced to zero by continuing the process of premultiplication with transformation matrices of the same type as the ones contained in $\mathcal{P}$, Eq. 177. Thus the fourth step may be the following:

$$\begin{bmatrix} 1 & 0 & -\frac{10}{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 2 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} = \begin{bmatrix} 1 & 3 & 0 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} \qquad [181]$$

The fifth step becomes

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{10}{13} \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 0 \\ 0 & -10 & -2 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} = \begin{bmatrix} 1 & 3 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} \qquad [182]$$

and the sixth and final step reads

$$\begin{bmatrix} 1 & \frac{3}{10} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & \frac{13}{5} \end{bmatrix} \qquad [183]$$

The resultant transformation matrix combining the last three steps becomes

$$\mathcal{Q} = \underbrace{\begin{bmatrix} 1 & \frac{3}{10} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Sixth step}} \times \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{10}{13} \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Fifth step}} \times \underbrace{\begin{bmatrix} 1 & 0 & -\frac{10}{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Fourth step}} = \begin{bmatrix} 1 & \frac{3}{10} & -\frac{7}{13} \\ 0 & 1 & \frac{10}{13} \\ 0 & 0 & 1 \end{bmatrix} \qquad [184]$$

which is, of course, different from the matrix 178, although not unlike this matrix in form. The complete transformation is accomplished by premultiplication of $\mathcal{A}$ by the resultant transformation matrix:

$$\mathcal{R} = \mathcal{Q} \times \mathcal{P} = \begin{bmatrix} 1 & \frac{3}{10} & -\frac{7}{13} \\ 0 & 1 & \frac{10}{13} \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{4}{13} & \frac{9.5}{13} & -\frac{7}{13} \\ -\frac{50}{13} & \frac{5}{13} & \frac{10}{13} \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix} \qquad [185]$$

In place of the transformation expressed by Eq. 176, one now has

$$\mathcal{R} \times \mathcal{A} = \mathcal{D} \qquad [186]$$

It should be clear that a similar procedure, involving postmultiplication only, can also be used to accomplish the same result. The reader may carry out the detailed steps as an exercise.

In the above numerical example, the rank of the given matrix is equal to its order; that is, the matrix is nonsingular. When the rank of $\mathcal{A}$ is less than its order, some of the diagonal elements in $\mathcal{D}$ are found to be zero, but the method of transformation given in Eq. 176 is still applicable.

From the definition of the rank of a square matrix (which is the same as the rank of its determinant) it is readily seen that if the rank of a matrix $\mathcal{A}$ of order n is $r = (n - p)$, then p of the diagonal elements in $\mathcal{D}$ become zero. It is then possible to arrange the remaining elements so that $\mathcal{D}$ has the appearance

$$\mathcal{D} = \begin{bmatrix} d_{11} & 0 & 0 & & \cdots & 0 \\ 0 & d_{22} & 0 & & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & d_{rr} & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & \cdots & & 0 \end{bmatrix} \qquad [187]$$

By a further transformation, the following form may be obtained:

$$\mathcal{E} = \begin{bmatrix} 1 & 0 & 0 & & \cdots & 0 \\ 0 & 1 & 0 & & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & & 0 \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & \cdots & & 0 \end{bmatrix} \left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} r \text{ rows} \qquad [188]$$

Since the matrix $\mathcal{D}$ is symmetric, it may be desirable to find $\mathcal{E}$ by a transformation of the form $\mathcal{E} = \mathcal{P}_t\mathcal{D}\mathcal{P}$. The matrix $\mathcal{P}$ will not be real, however, if any of the diagonal elements in $\mathcal{D}$ are negative. The form of $\mathcal{E}$ given by Eq. 188 may nevertheless be obtained using only real matrices either by pre- or postmultiplication alone, or by a transformation like Eq. 176 in which $\mathcal{P}$ and $\mathcal{Q}$ need not be specially related.

This diagonal matrix (in which the first r diagonal elements are unity and the rest zero) is referred to as the *canonical* form of the square matrix $\mathcal{A}$ of rank r. The diagonal and the canonical forms of a matrix obviously place its rank in evidence.

A square matrix whose rank is less than its order is said to be degenerate. If the rank is $r = (n - p)$, then p is spoken of as the *degree* of

degeneracy of the matrix or also as the *nullity*. The latter term is evidently suggested by the forms given by Eqs. 187 and 188.

When the matrix $\mathcal{A}$ is symmetrical, the matrix $\mathcal{P}$ which enters the transformation to the diagonal according to the process indicated in Eq. 176 can evidently be equal to the transpose of $\mathcal{Q}$,* so that this equation may be written

$$\mathcal{Q}_t \times \mathcal{A} \times \mathcal{Q} = \mathcal{D} \tag{189}$$

The matrix $\mathcal{Q}$ can have a variety of forms, depending upon the values of the diagonal elements appearing in $\mathcal{D}$. For certain values, called the *characteristic values* or also the *latent roots* of the matrix $\mathcal{A}$, the transformation matrix $\mathcal{Q}$ in Eq. 189 becomes *orthogonal*. This orthogonal transformation is discussed in detail in the immediately following chapters, where it is also shown that the corresponding values of the diagonal elements in $\mathcal{D}$ are the roots of the so-called *characteristic equation*,

$$\begin{vmatrix} (a_{11}-\lambda) & a_{12} & \cdots & a_{1n} \\ a_{21} & (a_{22}-\lambda) & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & (a_{nn}-\lambda) \end{vmatrix} = 0 \tag{190}$$

The left-hand side of this characteristic equation, which is formed by setting the determinant of the matrix $\mathcal{A} - \lambda\mathcal{U}$ equal to zero, is evidently a polynomial in λ of the nth degree. It is clear that the constant term in this equation (the term without λ) is equal to the determinant of $\mathcal{A}$. Moreover, the coefficient of the highest-power term is evidently $(-1)^n$. If, then, Eq. 190 is divided through by $(-1)^n$, and if the n roots are $\lambda_1, \lambda_2, \cdots \lambda_n$, it follows from the theory of algebraic equations that

$$A = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_n \tag{191}$$

The determinant of $\mathcal{D}$ is, of course, also given by the product of the roots $\lambda_1, \lambda_2, \cdots \lambda_n$, these being the diagonal elements in $\mathcal{D}$. This result checks with the fact that the determinant of an orthogonal matrix is ± 1 (see Eq. 86), so that the determinant of $\mathcal{D}$ in Eq. 189 is equal to the determinant of $\mathcal{A}$. One thus recognizes the invariance of the determinant of a matrix to an orthogonal transformation of the form of Eq. 189.

*It is significant that $\mathcal{P}$ is not *necessarily* the transpose of $\mathcal{Q}$ when $\mathcal{A}$ is symmetrical. Thus if the first operation upon the rows of $\mathcal{A}$ effects an interchange in any two rows, the matrix encountered at the next step is no longer symmetrical. Symmetry can, of course, be restored by subsequently performing a similar operation upon the corresponding columns, but such a step is not necessary, and if it is not taken the matrices $\mathcal{P}$ and $\mathcal{Q}$ ultimately obtained will obviously not be each other's transpose.

11. Additional methods for obtaining the inverse of a matrix

Having determined the matrices $\mathcal{P}$ and $\mathcal{Q}$ which accomplish the reduction of a matrix $\mathcal{A}$ to its equivalent diagonal form, according to Eq. 176, one may proceed to find the inverse of $\mathcal{A}$ by means of the following reasoning. The first step is to form the inverse of both sides of Eq. 176, making use of the property expressed by Eq. 100. This gives

$$\mathcal{Q}^{-1} \times \mathcal{A}^{-1} \times \mathcal{P}^{-1} = \mathcal{D}^{-1} \qquad [192]$$

Premultiplying by $\mathcal{Q}$ and postmultiplying by $\mathcal{P}$, one has

$$\mathcal{A}^{-1} = \mathcal{Q} \times \mathcal{D}^{-1} \times \mathcal{P} \qquad [193]$$

Inasmuch as the matrices $\mathcal{Q}$ and $\mathcal{P}$ are already known, and the determination of $\mathcal{D}^{-1}$ according to Eq. 67 is a relatively simple task, the formation of $\mathcal{A}^{-1}$ by means of the result expressed by Eq. 193 involves only a nominal amount of additional computation. For the numerical example given in the preceding article (see Eqs. 177, 178, and 179) the Eq. 193 yields

$$\mathcal{A}^{-1} = \begin{bmatrix} 1 & -3 & -\frac{7}{5} \\ 0 & 1 & -\frac{1}{5} \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{10} & 0 \\ 0 & 0 & \frac{5}{13} \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix}$$
$$= \frac{1}{13}\begin{bmatrix} -4 & 9.5 & -7 \\ 5 & -0.5 & -1 \\ 1 & -4 & 5 \end{bmatrix} \qquad [194]$$

Alternatively one may begin with Eq. 186 and form its inverse, which reads

$$\mathcal{A}^{-1} \times \mathcal{R}^{-1} = \mathcal{D}^{-1} \qquad [195]$$

Postmultiplication by $\mathcal{R}$ again gives the desired result, namely

$$\mathcal{A}^{-1} = \mathcal{D}^{-1} \times \mathcal{R} \qquad [196]$$

For the same numerical problem as discussed above (see Eq. 185) this reads

$$\mathcal{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{10} & 0 \\ 0 & 0 & \frac{5}{13} \end{bmatrix} \times \begin{bmatrix} -\frac{4}{13} & \frac{9.5}{13} & -\frac{7}{13} \\ -\frac{50}{13} & \frac{5}{13} & \frac{10}{13} \\ \frac{1}{5} & -\frac{4}{5} & 1 \end{bmatrix}$$
$$= \frac{1}{13}\begin{bmatrix} -4 & 9.5 & -7 \\ 5 & -0.5 & -1 \\ 1 & -4 & 5 \end{bmatrix} \qquad [197]$$

If $\mathcal{A}$ is reduced to the diagonal form by postmultiplication alone, that

is, if one has found a matrix $\mathcal{S}$ such that

$$\mathcal{A} \times \mathcal{S} = \mathcal{D} \tag{198}$$

then a similar manipulation as given above shows that

$$\mathcal{A}^{-1} = \mathcal{S} \times \mathcal{D}^{-1} \tag{199}$$

Another method of matrix inversion, which in its essential steps utilizes the reasoning involved in the formation of the matrix $\mathcal{P}$ in Eq. 177, is found to be exceptionally effective in numerical work.* The procedure is best described by considering the equivalent problem of solving a set of n simultaneous equations involving n unknowns, such as

$$\begin{matrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{matrix} \tag{200}$$

Any other set of n equations which are linear combinations of these (called an equivalent set of equations) clearly has the same solutions. Since only the values of the coefficients a_{sk} and those of the quantities $y_1 \cdots y_n$ are significant in the formation of such linear combinations, it is expedient to carry out the contemplated manipulations upon the matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & y_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & y_n \end{bmatrix} \tag{201}$$

which is the matrix $\mathcal{A}$ of the set of Eqs. 200 with the quantities $y_1 \cdots y_n$ included as an additional column. This resultant matrix involving n rows and $n + 1$ columns is referred to as the *augmented matrix* corresponding to the Eqs. 200.

In order to simplify the notation in the following discussion it is effective to write the symbols $a_{1,n+1}, a_{2,n+1}, \cdots a_{n,n+1}$, in place of y_1, $y_2, \cdots y_n$, so that the augmented matrix assumes the form

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \cdots & a_{2n} & a_{2,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & a_{n,n+1} \end{bmatrix} \tag{202}$$

The series of manipulations which are described presently, have for their objective the derivation of an equivalent set of equations having an augmented matrix in which all the elements of the principal diagonal

*This method is described in a paper by Prescott D. Crout entitled: "A Short Method for Evaluating Determinants and Solving Systems of Linear Equations with Real or Complex Coefficients," *A.I.E.E. Transactions*, Vol. 60, 1941, pp. 1235–1240.

are unity and all those below this diagonal are zero. It is readily appreciated that once this special set of equivalent equations is determined, the desired solutions may be written down by inspection.

The first step is to divide the elements of the first row in the matrix 202 by a_{11} so as to obtain

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n+1} \end{bmatrix} \qquad [203]$$

This step requires that a_{11} have a nonzero value. If this condition is not fulfilled it can always be met by a rearrangement of the original equations. The a_{21}-multiplied elements of the first row are now subtracted from the corresponding elements of the second row, giving

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ 0 & b_{22} & b_{23} & \cdots & b_{2,n+1} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n+1} \end{bmatrix} \qquad [204]$$

in which

$$\begin{aligned} b_{22} &= a_{22} - \frac{a_{21}a_{12}}{a_{11}} \\ b_{23} &= a_{23} - \frac{a_{21}a_{13}}{a_{11}} \\ &\cdots\cdots\cdots\cdots \\ b_{2,n+1} &= a_{2,n+1} - \frac{a_{21}a_{1,n+1}}{a_{11}} \end{aligned} \qquad [205]$$

Next the a_{31}-multiplied elements of the first row are subtracted from the corresponding elements of the third row. This leaves the augmented matrix in the form

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ 0 & b_{22} & b_{23} & \cdots & b_{2,n+1} \\ 0 & b_{32} & b_{33} & \cdots & b_{3,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n+1} \end{bmatrix} \qquad [206]$$

with

$$b_{32} = a_{32} - \frac{a_{31}a_{12}}{a_{11}}$$

$$b_{33} = a_{33} - \frac{a_{31}a_{13}}{a_{11}} \qquad [207]$$

$$b_{3,n+1} = a_{3,n+1} - \frac{a_{31}a_{1,n+1}}{a_{11}}$$

The procedure is continued until the matrix assumes the form

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ 0 & b_{22} & b_{23} & \cdots & b_{2,n+1} \\ 0 & b_{32} & b_{33} & \cdots & b_{3,n+1} \\ 0 & b_{n2} & b_{n3} & \cdots & b_{n,n+1} \end{bmatrix} \qquad [208]$$

A single formula for the coefficients b_{sk} is recognized to be given by

$$b_{sk} = a_{sk} - \frac{a_{s1}a_{1k}}{a_{11}} \qquad [209]$$

Now the elements of the second row are divided by b_{22}. (A rearrangement of the last $n - 1$ rows will be required if b_{22} happens to be zero.) One then has

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & & \frac{a_{1,n+1}}{a_{11}} \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \cdots & \frac{b_{2,n+1}}{b_{22}} \\ 0 & b_{32} & b_{33} & \cdots & b_{3,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & b_{n2} & b_{n3} & \cdots & b_{n,n+1} \end{bmatrix} \qquad [210]$$

The portion of this matrix exclusive of the first row and column is now dealt with in a manner which is identical with that just described for the transformation of the matrix 203 into the form of 208. In the first step the b_{32}-multiplied elements of the second row are subtracted from the corresponding ones of the third row. Next the b_{42}-multiplied elements of the second row are subtracted from the corresponding elements of the fourth row, and so on. This set of operations finally yields the matrix in the form

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{a_{14}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \frac{b_{24}}{b_{22}} & \cdots & \frac{b_{2,n+1}}{b_{22}} \\ 0 & 0 & c_{33} & c_{34} & \cdots & c_{3,n+1} \\ 0 & 0 & c_{43} & c_{44} & \cdots & c_{4,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & c_{n3} & c_{n4} & \cdots & c_{n,n+1} \end{bmatrix} \qquad [211]$$

in which the coefficients c_{sk} are given by the formula

$$c_{sk} = b_{sk} - \frac{b_{s2}b_{2k}}{b_{22}} \qquad [212]$$

The portion of this matrix involving the coefficients c_{sk} is now treated in the same manner as described for the original matrix 202, and as repeated in the manipulation of that portion of the matrix 208 involving the coefficients b_{sk}. The resulting matrix then assumes the form

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{a_{14}}{a_{11}} & \cdots & \frac{a_{1,n+1}}{a_{11}} \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \frac{b_{24}}{b_{22}} & \cdots & \frac{b_{2,n+1}}{b_{22}} \\ 0 & 0 & 1 & \frac{c_{34}}{c_{33}} & \cdots & \frac{c_{3,n+1}}{c_{33}} \\ 0 & 0 & 0 & d_{44} & \cdots & d_{4,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & d_{n4} & \cdots & d_{n,n+1} \end{bmatrix} \qquad [213]$$

with

$$d_{sk} = c_{sk} - \frac{c_{s3}c_{3k}}{c_{33}} \qquad [214]$$

The continuation of this process is obvious. In the case $n = 4$, for example, one arrives at the desired form

$$\begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{a_{14}}{a_{11}} & \frac{a_{15}}{a_{11}} \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \frac{b_{24}}{b_{22}} & \frac{b_{25}}{b_{22}} \\ 0 & 0 & 1 & \frac{c_{34}}{c_{33}} & \frac{c_{35}}{c_{33}} \\ 0 & 0 & 0 & 1 & \frac{d_{45}}{d_{44}} \end{bmatrix} \qquad [215]$$

The solutions to the equivalent equations having this augmented matrix are easily recognized. They are given by

$$
\begin{aligned}
x_4 &= \frac{d_{45}}{d_{44}} \\
x_3 &= \frac{c_{35}}{c_{33}} - \frac{c_{34}}{c_{33}} x_4 \\
x_2 &= \frac{b_{25}}{b_{22}} - \frac{b_{24}}{b_{22}} x_4 - \frac{b_{23}}{b_{22}} x_3 \\
x_1 &= \frac{a_{15}}{a_{11}} - \frac{a_{14}}{a_{11}} x_4 - \frac{a_{13}}{a_{11}} x_3 - \frac{a_{12}}{a_{11}} x_2
\end{aligned}
\qquad [216]
$$

In order to systematize the computational procedure it is expedient to record the following "auxiliary matrix":

$$
\begin{matrix}
a_{11} & \dfrac{a_{12}}{a_{11}} & \dfrac{a_{13}}{a_{11}} & \dfrac{a_{14}}{a_{11}} & \dfrac{a_{15}}{a_{11}} \\
a_{21} & b_{22} & \dfrac{b_{23}}{b_{22}} & \dfrac{b_{24}}{b_{22}} & \dfrac{b_{25}}{b_{22}} \\
a_{31} & b_{32} & c_{33} & \dfrac{c_{34}}{c_{33}} & \dfrac{c_{35}}{c_{33}} \\
a_{41} & b_{42} & c_{43} & d_{44} & \dfrac{d_{45}}{d_{44}}
\end{matrix}
\qquad [217]
$$

From the recursion formulas 209, 212, 214 which may be rewritten in the form

$$
\begin{aligned}
b_{sk} &= a_{sk} - \frac{a_{s1}a_{1k}}{a_{11}} \\
c_{sk} &= a_{sk} - \frac{a_{s1}a_{1k}}{a_{11}} - \frac{b_{s2}b_{2k}}{b_{22}} \\
d_{sk} &= a_{sk} - \frac{a_{s1}a_{1k}}{a_{11}} - \frac{b_{s2}b_{2k}}{b_{22}} - \frac{c_{s3}c_{3k}}{c_{33}}
\end{aligned}
\qquad [218]
$$

it becomes clear that the elements in the auxiliary matrix can be calculated and recorded in alphabetical sequence, with the recording of the elements of any column preceding that of the elements of the corresponding row.

Although only the elements above the principal diagonal are needed in the final computation of the unknowns, as is shown by the Eqs. 216, it is necessary to calculate and record also the elements on and below the principal diagonal since their values are needed in the sequence of computations which determine the auxiliary matrix as a whole. This fact is readily appreciated from an examination of the recursion formulas 218, or better still through an application of them to a numerical example.

Since only those results are recorded which are needed in subsequent calculations, the computational scheme is orderly and compact. It is also significant to mention that the calculation of any one of the elements of the auxiliary matrix, in the sequence described above, may be accomplished through a single continuous operation on a modern computing machine. Thus, from the standpoint of minimizing the required operations, the present procedure for solving simultaneous equations is excellently adapted to the available computing facilities.

Regarding the problem of inverting the matrix of Eqs. 200 it is helpful to make the following preliminary observations. Suppose, in this system of equations, that all but one of the quantities $y_1 \cdots y_n$ are zero, and that the nonzero one, which may be any y_k, has the value unity. The particular solution corresponding to this choice of y-values may conveniently be denoted by $x_{1k}, x_{2k}, \cdots x_{nk}$. Since the index k may have any value from 1 to n, there are n such sets of particular solutions. It should now be recognized that the set of x-values for $k = 1$, when arranged in a column in which the order of numerical subscripts reads from top to bottom, represents the first column in the desired inverse matrix; the set for $k = 2$ represents the second column in the inverse matrix, and so forth. Thus the complete array of n^2 quantities x_{sk} are the elements of the inverse matrix, with the customary denotation of the indexes s and k.

The process of determining the inverse matrix is thus seen to require the evaluation of n sets of simultaneous solutions. This process does not, however, require n times the computational labor involved in obtaining a single solution inasmuch as only the last column in the auxiliary matrix differs in the procedure for obtaining each solution. The work may be arranged so that all the n solutions are evaluated in one continuous sequence of operations by recording an auxiliary matrix with n varieties of " last " columns written side by side. The details of such a procedure are best illustrated by means of a numerical example.

Choosing the same matrix as is used in the numerical illustrations of the preceding article, one begins by considering the following augmented matrix

$$\begin{bmatrix} 1 & 3 & 2 & 1 & 0 & 0 \\ 4 & 2 & 6 & 0 & 1 & 0 \\ 3 & 1 & 7 & 0 & 0 & 1 \end{bmatrix} \qquad [219]$$

in which the last three columns represent three varieties of the column of y-values as given in the matrix 201. The first of these three columns represents the values $y_1 = 1$, $y_2 = y_3 = 0$; for the second of the last three columns, $y_2 = 1$, $y_1 = y_3 = 0$; and for the third, $y_3 = 1$, $y_1 = y_2 = 0$. Regarding all the elements of this augmented matrix as coefficients a_{ik} in the recursion formulas 218, one obtains the auxiliary matrix

$$\begin{matrix} 1 & 3 & 2 & 1 & 0 & 0 \\ 4 & -10 & \frac{2}{10} & \frac{4}{10} & -\frac{1}{10} & 0 \\ 3 & -8 & \frac{13}{5} & \frac{1}{13} & -\frac{4}{13} & \frac{5}{13} \end{matrix} \qquad [220]$$

Only the element values above and to the right of the elements on the principal diagonal (beyond the dotted line) are now used to determine the values of the quantities x_{ik}. The fourth, fifth, and sixth columns are regarded as "last" columns in computing values of x_{ik} for $k = 1, 2, 3$ respectively. For $k = 1$, one finds

$$\begin{aligned} x_{31} &= \frac{1}{13} \\ x_{21} &= 0.4 - 0.2x_{31} = \frac{5}{13} \\ x_{11} &= 1 - 2x_{31} - 3x_{21} = -\frac{4}{13} \end{aligned} \qquad [221]$$

For $k = 2$,

$$\begin{aligned} x_{32} &= -\frac{}{13} \\ x_{22} &= -\frac{1}{10} - 0.2x_{32} = -\frac{0.5}{13} \\ x_{12} &= 0 - 2x_{32} - 3x_{22} = \frac{9.5}{13} \end{aligned} \qquad [222]$$

and for $k = 3$,

$$\begin{aligned} x_{33} &= \frac{}{13} \\ x_{23} &= 0 - 0.2x_{33} = -\frac{1}{13} \\ x_{13} &= 0 - 2x_{33} - 3x_{23} = -\frac{7}{13} \end{aligned} \qquad [223]$$

Hence the desired inverse matrix is

$$\mathfrak{A}^{-1} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} = \frac{1}{13}\begin{bmatrix} -4 & 9.5 & -7 \\ 5 & -0.5 & -1 \\ 1 & -4 & 5 \end{bmatrix} \qquad [224]$$

which agrees with the results given by Eqs. 194 and 197.

PROBLEMS

1. Write the following as a matrix equation:

$$P = e_1 i_1 + e_2 i_2 + e_3 i_3 + e_4 i_4$$

For example,

$$51 = 2 \times 5 - 3 \times 7 + 6 \times 9 + 4 \times 2$$

2. Write the four scalar equations:

$$x_1 = y_1 \qquad x_2 = y_2 \qquad x_3 = y_3 \qquad x_4 = y_4$$

as a single matrix equation in at least four different ways.

3. Express as matrix equations the relations:

$$a_{sk} = \frac{\partial^2 f}{\partial x_s \partial x_k} \qquad \text{and} \qquad a_{sk} = \frac{\partial A_s}{\partial x_k}$$

for $s = 1, 2, 3$ and $k = 1, 2, 3$.

4. Evaluate:

$$\begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix} + \begin{bmatrix} -2 & 3 \\ 1 & 9 \end{bmatrix} - \begin{bmatrix} 6 & 4 \\ -2 & 7 \end{bmatrix}$$

and

$$6\begin{bmatrix} 1 & 6 & 2 \\ 4 & 3 & -5 \\ 7 & -9 & 1 \end{bmatrix} - 4\begin{bmatrix} 2 & 9 & -6 \\ 4 & -7 & 10 \\ 20 & 8 & -5 \end{bmatrix}$$

5. Carry out the following triple product in two ways and note the difference in total numbers of multiplications and additions:

$$\begin{bmatrix} 3 & 2 & -1 & 6 \\ 5 & -7 & 9 & 8 \\ -4 & 10 & 6 & 2 \end{bmatrix} \times \begin{bmatrix} 1 & 3 \\ 6 & -9 \\ -4 & 7 \\ 2 & 11 \end{bmatrix} \times \begin{bmatrix} 6 & 4 \\ -2 & 1 \end{bmatrix}$$

6. Evaluate:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 7 & 2 & -10 \\ -8 & 4 & 3 \\ 1 & -6 & 7 \end{bmatrix}$$

and

$$\begin{bmatrix} 7 & 2 & -10 \\ -8 & 4 & 3 \\ 1 & -6 & 7 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

7. Given:

$$a = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \qquad b = \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \qquad c = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Show that these are orthogonal matrices. Also form the products ab, ac, bc, and abc, and show that they too are orthogonal. Note that $ab \neq ba$, $ac \neq ca$, etc.

8. In the evaluation of the following multiple product:

$$\begin{bmatrix} 2 \\ -3 \\ 5 \end{bmatrix} \times [4 \quad 6 \quad 3] \times \begin{bmatrix} 1 \\ -2 \\ 5 \end{bmatrix} \times [3 \quad 2]$$

determine that association which requires the least number of multiplications and additions.

9. Evaluate:

$$\begin{bmatrix} 2 & -6 & 4 \\ 3 & 9 & 7 \\ 1 & 5 & 8 \end{bmatrix} \times \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \times \begin{bmatrix} 2 & -6 & 4 \\ 3 & 9 & 7 \\ 1 & 5 & 8 \end{bmatrix}$$

10. Determine the most convenient ways of evaluating the following multiple products:

$$\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 6 \\ -5 & 7 \end{bmatrix} \times \begin{bmatrix} 6 & 1 \\ 4 & -3 \end{bmatrix} \times \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

$$[5 \quad 7] \times \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 6 \\ -5 & 7 \end{bmatrix} \times \begin{bmatrix} 6 & 1 \\ 4 & -3 \end{bmatrix}$$

11. Evaluate:

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}^2 \quad \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}^3 \quad \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}^{-1} \quad \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}^{-2}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}$$

12. Using the indicated partitioning, find the inverse of the following product:

$$\left[\begin{array}{cc:cc} 2 & 6 & 0 & 0 \\ 1 & 7 & 0 & 0 \\ \hdashline 0 & 0 & 5 & 4 \\ 0 & 0 & 3 & 1 \end{array}\right] \times \left[\begin{array}{cc:cc} -1 & 2 & 0 & 0 \\ 3 & -4 & 0 & 0 \\ \hdashline 0 & 0 & 5 & -2 \\ 0 & 0 & 7 & -6 \end{array}\right]$$

13. Given the nonsingular matrix:

$$\mathcal{A} \quad \begin{bmatrix} 1 & 0 & -2 \\ 4 & 3 & 5 \\ 0 & -6 & 8 \end{bmatrix}$$

find the adjoint $\mathcal{A}^a$ and evaluate the products $\mathcal{A} \times \mathcal{A}^a$ and $\mathcal{A}^a \times \mathcal{A}$. Repeat with the matrix:

$$\mathcal{A} = \begin{bmatrix} 1 & 3 \\ 2 & -5 \\ -4 & 6 \end{bmatrix} \times \begin{bmatrix} 4 & -2 & 3 \\ 1 & 0 & 1 \end{bmatrix}$$

Discuss the results, showing why their particular forms should be expected.

14. Using the matrices $\mathcal{A}$ of the preceding problem, determine matrices $\mathcal{P}$ and $\mathcal{Q}$ which will yield diagonal forms in each of the following applicable relations:

$$\text{(i) } \mathcal{P} \times \mathcal{A} \times \mathcal{Q} = \mathcal{D} \qquad \text{(ii) } \mathcal{P} \times \mathcal{A} = \mathcal{D} \qquad \text{(iii) } \mathcal{A} \times \mathcal{Q} = \mathcal{D}$$

In terms of these results, find the inverse of the nonsingular matrix of Prob. 13.

15. Given:

$$\mathcal{A} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 4 & -3 & 2 & 1 \\ 6 & 2 & 1 & 4 \\ -3 & 0 & 5 & 7 \end{bmatrix}$$

find $\mathcal{A}^{-1}$, using the method of partitioning discussed in Art. 7, and again by the method involving the "auxiliary matrix" as discussed in Art. 11.

16. Prove directly from the definitions of the sum and product of two matrices that

$$\mathcal{A} \times (\mathcal{B} \times \mathcal{C}) = (\mathcal{A} \times \mathcal{B}) \times \mathcal{C}$$

and

$$\mathcal{A} \times (\mathcal{B} + \mathcal{C}) = (\mathcal{A} \times \mathcal{B}) + (\mathcal{A} \times \mathcal{C})$$

17. Compute the third power of the matrix

$$\mathcal{A} = \begin{bmatrix} 1 & -1 & 2 & 1 \\ -1 & 3 & 0 & -3 \\ 2 & 0 & 9 & -6 \\ 1 & -3 & -6 & 19 \end{bmatrix}$$

and check the value 13824 of the determinant of the resultant matrix.

18. If $\mathcal{A}$ is a nonsingular matrix, and the expression $\mathcal{A}^{-n}$ is interpreted as being equivalent to $(\mathcal{A}^{-1})^n$, show that $\mathcal{A}^{-n}$ is unique and, in particular, that

$$\mathcal{A}^0 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

19. The row matrix $[1 \quad y_2 \quad y_3 \cdots y_n]$ may be regarded as equivalent to a square matrix having these elements in its first row and zeros in the remaining rows. With this interpretation of the row matrix, show that

$$[1 \quad y_2 \quad y_3 \cdots y_n]^m = [1 \quad y_2 \quad y_3 \cdots y_n]$$

and more generally that

$$[y_1 \quad y_2 \cdots y_n]^m = y_1^{m-1}[y_1 \quad y_2 \cdots y_n]$$

Analogously, find for the column matrix that

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}^m \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

20. The expression

$$P(\mathcal{X}) = \mathcal{A}_0 + \mathcal{A}_1\mathcal{X} + \mathcal{A}_2\mathcal{X}^2 + \cdots + \mathcal{A}_n\mathcal{X}^n$$

in which $\mathcal{A}_0, \mathcal{A}_1, \cdots \mathcal{A}_n$ are matrix coefficients and $\mathcal{X}$ is a matrix playing the role of a variable, may be regarded as a matrix polynomial. A less general form of matric polynomial is

$$Q(\mathcal{X}) = p_0\mathcal{U} + p_1\mathcal{X} + p_2\mathcal{X}^2 + \cdots + p_n\mathcal{X}^n$$

in which $p_0, p_1, \cdots p_n$ are scalar coefficients and $\mathcal{U}$ is a unit matrix having the same order as $\mathcal{X}$.

(a) Show that the equation

$$\mathcal{A}\mathcal{X} + \mathcal{B} = 0$$

in which $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{B}$ are matrices, has one solution only if $\mathcal{A}$ is nonsingular.

(b) In the equation

$$\mathcal{A}_n\mathcal{X}^n + \mathcal{A}_{n-1}\mathcal{X}^{n-1} + \cdots + \mathcal{A}_1\mathcal{X} + \mathcal{A}_0 = 0$$

let

$$\mathcal{X} = [x_1 \quad x_2 \cdots x_n]$$

the coefficients $\mathcal{A}$ being matrices. Show that a solution is given by the expression

$$\mathcal{X} = -(\mathcal{A}_n x_1^{n-1} + \mathcal{A}_{n-1}x_1^{n-2} + \cdots + \mathcal{A}_1)^{-1}\mathcal{A}_0$$

provided the matrix in the parenthesis is nonsingular.

(c) Consider the polynomial

$$\mathcal{Y} = \mathcal{X}^2 + \mathcal{B}\mathcal{X} + \mathcal{C}$$

with matrix coefficients, and

$$\mathcal{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

Show that a solution to the equation $\mathcal{Y} = 0$ (null matrix) exists only if the following conditions are fulfilled

$$\begin{aligned} x_{11}^2 + x_{12}x_{21} + b_{11}x_{11} + b_{12}x_{21} &= -C_{11} \\ (x_{11} + x_{22} + b_{11})x_{12} + b_{12}x_{22} &= -C_{12} \\ (x_{11} + x_{22} + b_{22})x_{21} + b_{21}x_{11} &= -C_{21} \\ x_{22}^2 + x_{21}x_{12} + b_{21}x_{12} + b_{22}x_{22} &= -C_{22} \end{aligned}$$

and that the elements $x_{11}, x_{22}, x_{12}, x_{21}$ are independent only if

$$J \quad \begin{matrix} 2x_{11} + b_{11} & x_{21} & x_{12} + b_{12} & 0 \\ x_{12} & x_{11} + x_{22} + b_{11} & 0 & x_{12} + b_{12} \\ x_{21} + b_{21} & 0 & x_{11} + x_{22} + b_{22} & x_{21} \\ 0 & x_{21} + b_{21} & x_{12} & 2x_{22} + b_{22} \end{matrix} \quad \neq 0$$

21. $\mathcal{A}$ and $\mathcal{B}$ are two square matrices of like order whose product is zero, that is, $\mathcal{A} \times \mathcal{B} = 0$ in which 0 is a null matrix of the same order as $\mathcal{A}$ or $\mathcal{B}$. Let $\mathcal{A}$ be given and $\mathcal{B}$ unknown. Write the complete set of equations whose solution (if it exists) yields the elements b_{ik} in terms of the elements a_{ik}. If $\mathcal{A}$ has the order n and the rank n, what is the solution? Discuss the possibility and form of solutions if $\mathcal{A}$ has the rank $n - 1$, and more generally if $\mathcal{A}$ has the rank $n - p$.

Solve the specific equations

$$\begin{bmatrix} 1 & 5 & 1 \\ 3 & 15 & 0 \\ 4 & 20 & 1 \end{bmatrix} \times \mathfrak{B} = 0 \qquad \text{and} \qquad \mathfrak{B} \times \begin{bmatrix} 1 & 5 & 1 \\ 3 & 15 & 0 \\ 4 & 20 & 1 \end{bmatrix} = 0$$

22. Given the equation

$$\mathfrak{X}^2 + 3\mathfrak{X} - 10\mathfrak{U} = 0$$

in which $\mathfrak{X}$ is a square matrix and $\mathfrak{U}$ is the unit matrix of like order. Find solutions of the form $x_\nu\mathfrak{U}$ in which x_ν is a scalar and show that one may write the matrix polynomial in factored form

$$(\mathfrak{X} - x_1\mathfrak{U}) \times (\mathfrak{X} - x_2\mathfrak{U}) = \mathfrak{X}^2 + 3\mathfrak{X} - 10\mathfrak{U}$$

Check this result for the particular value

$$\mathfrak{X} \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

23. If $\mathfrak{C}$ is a matrix with complex elements $c_{ik} = a_{ik} + jb_{ik}$ and if $\overline{\mathfrak{C}}$ has the conjugate elements, show that although $\mathfrak{C} + \overline{\mathfrak{C}}$ is real and $\mathfrak{C} - \overline{\mathfrak{C}}$ is imaginary, it does not follow in general that $\mathfrak{C} \times \overline{\mathfrak{C}}$ is real nor that $\mathfrak{C} \times \overline{\mathfrak{C}} = \overline{\mathfrak{C}} \times \mathfrak{C}$.

CHAPTER III

Linear Transformations

1. Vector sets

The primary object of this chapter is to offer a means of visualization for the purely algebraic reasoning underlying the essential ideas presented in Chs. I and II. This object is accomplished by giving a geometrical interpretation to the linear transformation and the matrix which characterizes it. There are, of course, a variety of possible forms which such a geometrical interpretation may take. The point of view given here is chosen for its utility in the particular applications discussed in the reference volumes.

In order to keep the discussion fairly general initially, the matrix is assumed to have m rows and n columns instead of being square. It is written

$$\mathcal{A} = \tag{1}$$

If the elements of each row are considered to be the components of a space vector with reference to a chosen rectangular Cartesian co-ordinate system, the matrix assumes an easily visualized geometrical characterization. It represents a set of vectors all emanating from a common origin so as to form a cluster. Since the number of components of each vector is in general greater than three, the process of visualization lacks the physical clarity which ordinarily accompanies the conception of a space vector. A purely mathematical extension of the idea of ordinary space to the conception of a space of many dimensions involves, however, only a passing mental hazard.

The co-ordinate axes in this many-dimensional space are for convenience numbered 1, 2, 3, $\cdots$ etc., instead of being lettered x, y, z, as in the usual three-dimensional case. Thus the elements a_{11}, a_{12}, $\cdots$ a_{1n} of the first row of $\mathcal{A}$, for example, are looked upon as the components (projections) of a vector a_1 with reference to the rectangular co-ordinate axes 1, 2, $\cdots$ n in an n-dimensional space. In like manner, the elements a_{21}, a_{22}, $\cdots$ a_{2n} are looked upon as the components of a vector a_2. Finally, the elements of the mth row define a vector a_m. The cluster of vectors a_1, a_2, $\cdots$ a_m is spoken of as the *vector set* of $\mathcal{A}$.

It is possible for the matrix $\mathcal{A}$ to be characterized by an alternative

vector set, namely the n vectors $a_1', a_2', \cdots a_n'$ whose components are the elements taken by *columns* instead of by rows. In this case, the space occupied by the vectors is m-dimensional, because each vector now has m components. This cluster of vectors is called the transposed vector set of $\mathcal{A}$, since it is the vector set of the transpose of $\mathcal{A}$.

2. Linear dependence and independence; the rank of a vector set

If one or more of the vectors in a set can be expressed as a linear vector addition of the remaining vectors multiplied by suitable positive or negative numerical factors, the set is said to be *linearly dependent.* If such an expression is not possible, the vectors are *linearly independent.* For a set of m vectors the possible existence of either condition may be expressed mathematically by the statement that if one can *not* find m numbers $\gamma_1, \gamma_2, \cdots \gamma_m$ (excluding the choice $\gamma_1 = \gamma_2 = \cdots = \gamma_m = 0$) for which the following vector equation holds

$$\gamma_1 a_1 + \gamma_2 a_2 + \cdots + \gamma_m a_m = 0 \tag{2}$$

the vectors form an independent (otherwise a dependent) set.

In three-dimensional space, any three vectors which do not lie in a plane form an independent set. If the three vectors do lie in a plane, one or more of them evidently can be expressed as a vector sum of the others multiplied by suitable factors. If in this same space the number of vectors is greater than three, the set must necessarily be dependent. In order for the vectors to form an independent set, it is necessary that their number shall not exceed three, although this condition alone is clearly not sufficient.

In an n-dimensional space, the largest number of independent vectors is n, but a dependent set may contain any number of vectors. The number of available dimensions in such a space is n. The vectors occupying this space may have such relative orientations that fewer than n dimensions are actually consumed or utilized. In three-dimensional space, for example, a set of vectors which lie in a plane consumes only two dimensions, and if they lie in a line they utilize only one of the available three dimensions.

It is clear that if the number of dimensions utilized is less than the number of vectors in the set, the vectors are linearly dependent; but if the number of utilized dimensions equals the number of vectors (it obviously cannot be greater than the number of vectors in the set), the vectors are linearly independent. If the number of vectors is equal to or less than the number of available dimensions, both cases can occur; but if the number

of vectors is larger than the number of available dimensions, the only possibility is for the vector set to be a dependent one.

The number of dimensions actually consumed by a vector set is by definition equal to the *rank* of that set. The vector set of $\mathfrak{a}$, Eq. 1, consists of m vectors in an n-dimensional space. The rank of this set can at most be equal to n. If $m > n$, the vector set of $\mathfrak{a}$ must be dependent, for the rank is then necessarily less than the number of vectors. In order to determine the rank, one may begin by considering consecutively all groups of n vectors which can be selected from the given set (the number of such groups equals the number of combinations of m things taken n at a time) and examining them to determine whether a linear relation of the form given by Eq. 2 can be found to exist. If one or more groups can be found for which such a relation does *not* exist, the rank is n; if it exists for *all* groups, the rank is less than n, and one must proceed to investigate in the same fashion all groups of $(n-1)$ vectors which can be selected from the given set. The largest number of independent vectors which can eventually be selected in this manner is equal to the rank of the vector set.

In the light of the discussion in Art. 10, Ch. I, regarding the rank of a determinant, the procedure just described may be seen to be equivalent to the following statements. From the matrix $\mathfrak{a}$ (with $m > n$), all the n-rowed determinants are selected. There are as many of these as there are m things taken n at a time, and they correspond to the groups of n vectors previously mentioned. The highest rank to be found among these determinants is the rank of a vector set of $\mathfrak{a}$.

In order to appreciate the truth of the latter assertion, it is necessary first to observe that the independence of $(n - 1)$ vectors in an n-dimensional space may be established by consideration of all the " projections " of these vectors onto the n co-ordinate "planes," comprising $(n - 1)$ dimensions each. Such a " projection " is carried out by simply allowing the components of all the vectors along one particular axis of the space to go to zero, or be disregarded. Consideration of the problem of two vectors in three dimensions will lead to the recognition that a necessary and sufficient condition for independence of the $(n - 1)$ vectors is that at least one of these n projected sets should be independent. Similarly, to insure dependence of the set, all the n projected sets must be dependent. In an identical manner the independence of $(n - p)$ vectors in an n-dimensional space is determined by the independence of at least one of the group of all possible $(n - p)$-dimensional "projections" of the set, formed by striking out p components at a time from all the vectors. If, and only if, all such projected sets are dependent, the set itself is dependent. It is then pointed out that all possible groups of $(n - p)$ vectors which can be formed from the original group of m may be regarded as

selected by choosing all possible groups of $(n - p)$ vectors from each of the groups of n vectors originally selected. The process of ascertaining the highest rank among the n determinants first chosen is then seen to be exactly equivalent to finding the largest number of independent vectors in the set by a method of successive projections.

In the transposed vector set of $\mathfrak{A}$ the number of vectors is n and the number of available dimensions is m. Still assuming $m > n$, the rank can at most be n, and is equal to the largest number of independent vectors which can be formed from the given set. The dependence or independence of a selected group of vectors is ascertained through noting whether or not a relation of the form given by Eq. 2 exists. This process of investigation is, however, equivalent to determining the highest rank to be found among all determinants formed from the transpose of $\mathfrak{A}$ by selecting n columns in all possible combinations. These determinants are simply the transposed ones of those referred to in the preceding paragraph. Since the rank of a determinant is the same as that of its transpose, it therefore follows that *the vector set of the transpose of $\mathfrak{A}$ has the same rank as the vector set of $\mathfrak{A}$.*

The same conclusion holds when $m < n$ and $m = n$. When $m \neq n$, the rank of the vector set or its transpose is at most equal to the smaller of the numbers m and n. When $m = n$, which is the case of greatest practical importance, the definition of the rank of the vector sets agrees with that already given for the rank of a square matrix and its determinant.*

3. Vector significance of a linear transformation

The ideas developed in the preceding articles suggest a useful geometrical interpretation for the linear transformation

$$\left.\begin{array}{l} a_{11}x_1 + \cdots + a_{1n}x_n = y_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n = y_n \end{array}\right\} \qquad [3]$$

with the square matrix

$$\mathfrak{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \qquad [4]$$

In the n-dimensional space occupied by the vector set of $\mathfrak{A}$, two additional vectors x and y, with components $x_1, x_2, \cdots x_n$ and $y_1, y_2, \cdots y_n$ respectively, are defined. The left-hand sides of the equations in the set 3

*In connection with the discussion of this article it may be of interest also to refer to Art. 6, Ch. IV, for an alternative method for establishing the linear dependence or independence of a given vector set.

are then recognized to be the scalar products of the vectors $a_1, a_2, \cdots a_n$ respectively and the vector x.

The scalar product (cf. Art. 2, Ch. V) is expressible either as the sum of the products of corresponding components of two vectors or as the product of their lengths and the cosine of the angle between them. If the given vectors are at right angles to each other, their scalar product is evidently zero.

With the scalar product denoted by a dot placed between the symbols for the vectors (Gibbs's notation), Eqs. 3 may be written in the alternative compact form

$$\left.\begin{array}{l} a_1 \cdot x = y_1 \\ a_2 \cdot x = y_2 \\ \cdots\cdots\cdots \\ a_n \cdot x = y_n \end{array}\right\} \tag{5}$$

These equations are said to transform a vector x into a vector y, the components of the latter being the scalar products of the vector x with those in the vector set of $\mathfrak{A}$.

Any vector x with given length and direction is transformed into another vector y by means of Eqs. 5, the mechanism of the transformation being determined by the cluster of vectors $a_1, a_2, \cdots a_n$. As the length and direction of the vector x are varied at will, the length and direction of the vector y vary in a corresponding manner. The vectors x and y may be visualized as rods emanating from a box which contains the mechanism of the transformation as characterized by the matrix $\mathfrak{A}$ or its vector set. As the rod representing x is pulled out or pushed in and its direction changed, the mechanism in the box causes the rod representing y to lengthen or shorten and change its direction in a corresponding manner.

For the moment, the mechanism in the box may be assumed to be so constituted that each length and direction of the rod representing x uniquely determines a length and direction for the rod representing y. This is evidently the case if the matrix $\mathfrak{A}$ is nonsingular. Then the transformation is reversible; that is, the rod representing y may be moved about at will, thus causing the mechanism in the box to produce corresponding lengths and directions for the rod x. A given point in space for the tip of the rod x determines one (and only one) point for the tip of the rod y, the corresponding points being independent of whether the rod x is pushed and the rod y caused to follow or vice versa.*

*In particular it may be of interest to note that if the vector x is moved about in a plane, the vector y remains in a plane also; and if the tip of the x-vector follows a straight line, the tip of the y-vector likewise follows a straight line. These matters are treated in greater detail in the discussion immediately following Eq. 168 in Art. 10.

Viewed in this mechanical manner, the transformation 5 and its inverse involve the same mechanism. Algebraically, however, the inverse transformation

$$\left.\begin{array}{l} b_{11}y_1 + \cdots + b_{1n}y_n = x_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ b_{n1}y_1 + \cdots + b_{nn}y_n = x_n \end{array}\right\} \qquad [6]$$

written in vector form

$$\left.\begin{array}{l} b_1 \cdot y = x_1 \\ b_2 \cdot y = x_2 \\ \cdots\cdots\cdots \\ b_n \cdot y = x_n \end{array}\right\} \qquad [7]$$

involves the vectors $b_1, b_2, \cdots b_n$, which are the vector set of the inverse matrix

$$\mathfrak{A}^{-1} = \mathfrak{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \qquad [8]$$

Since $\mathfrak{B} \times \mathfrak{A}$ equals the unit matrix, and is formed through multiplying the rows of $\mathfrak{B}$ by the columns of $\mathfrak{A}$, it follows that the scalar product

$$b_i \cdot a_k{}^t = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \qquad [9]$$

The vectors $a_1{}^t, a_2{}^t, \cdots a_n{}^t$ are the transposed vector set of $\mathfrak{A}$. More specifically, the relations 9 state that b_1, for example, stands at right angles to the vectors $a_2{}^t, a_3{}^t, \cdots a_n{}^t$, whereas its scalar product with $a_1{}^t$ is unity; b_2 stands at right angles to the vectors $a_1{}^t, a_3{}^t, \cdots a_n{}^t$, and its scalar product with $a_2{}^t$ is unity, and so forth. This situation offers an interesting geometrical view of the relation between a nonsingular matrix and its inverse.

If the square matrix $\mathfrak{A}$ is symmetrical,

$$\left.\begin{array}{l} a_1{}^t = a_1 \\ a_2{}^t = a_2 \\ \cdots\cdots \\ a_n{}^t = a_n \end{array}\right\} \qquad [10]$$

Then the orthogonal relationships just described hold directly for the vector set of $\mathfrak{A}$ and the vector set of its inverse.

4. Orthogonal co-ordinate transformations; orthogonal vector sets

It sometimes becomes necessary in dealing with physical problems to transform from one co-ordinate system to another. With respect to a

given reference system whose rectangular axes are numbered 1, 2, $\cdots n$, the co-ordinates of a particular point may be denoted correspondingly by the quantities $x_1, x_2, \cdots x_n$. In a second rectangular reference system whose origin coincides with that of the given one, the axes which are labeled $1', 2', \cdots n'$ have different angular orientations with respect to the axes of the original system. This second reference system may be thought of as obtained by simply rotating the axes of the given system, keeping the origin fixed and maintaining the mutual orthogonality of the axes.

The point in space whose projections on the original set of axes are the quantities $x_1, x_2, \cdots x_n$ has projections on the axes of the second reference system which are conveniently denoted by $x'_1, x'_2, \cdots x'_n$. These are the co-ordinates of the same point with respect to the second reference system. The process of expressing the quantities $x'_1, x'_2, \cdots x'_n$ in terms of $x_1, x_2, \cdots x_n$ is spoken of as a *co-ordinate transformation.*

In order for such a co-ordinate transformation to be carried out, the relative orientations of the axes in the original and the new reference systems must, of course, be known. Algebraically the directions of the new axes with respect to the old ones are expressed by quantities called *direction cosines.* If the cosines of the angles between axis $1'$ and the axes $1, 2, \cdots n$ are denoted respectively by the coefficients $o_{11}, o_{12}, \cdots o_{1n}$, the cosines of the angles between axis $2'$ and the axes $1, 2, \cdots n$ respectively by $o_{21}, o_{22}, \cdots o_{2n}$, and so forth, the array of coefficients in the matrix

$$\mathfrak{O} = \begin{bmatrix} o_{11} & o_{12} & \cdots & o_{1n} \\ o_{21} & o_{22} & \cdots & o_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ o_{n1} & o_{n2} & \cdots & o_{nn} \end{bmatrix} \qquad [11]$$

are the complete set of direction cosines of the new axes with respect to the old ones.

The desired co-ordinate transformation is then expressed by the linear set of equations

$$\left.\begin{array}{l} o_{11}x_1 + o_{12}x_2 + \cdots + o_{1n}x_n = x'_1 \\ o_{21}x_1 + o_{22}x_2 + \cdots + o_{2n}x_n = x'_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ o_{n1}x_1 + o_{n2}x_2 + \cdots + o_{nn}x_n = x'_n \end{array}\right\} \qquad [12]$$

In order to appreciate the correctness of this result with regard to its general form, and also to clarify the geometrical relations involved, the reader may carry through the derivation of this transformation for the simple two-dimensional case.

In the language of the vector interpretation given in the preceding

article, Eqs. 12 transform a vector x with the components $x_1, x_2, \cdots x_n$ into a vector x' with the components $x'_1, x'_2, \cdots x'_n$. Since the origin for both co-ordinate systems is the same and only one point in space is involved, x and x' are the same vector. It has a different direction with respect to the new reference axes from that which it has with respect to the old axes, but its length must certainly be the same in both reference systems. This fact means that

$$x_1^2 + x_2^2 + \cdots x_n^2 = x'^2_1 + x'^2_2 + \cdots + x'^2_n \tag{13}$$

The right-hand side of this equation may be obtained through multiplying Eqs. 12 respectively by $x'_1, x'_2, \cdots x'_n$ and adding. This process leads to a double summation. In order to abbreviate the writing of the result, it is effective to use the summation sign. Thus Eq. 13 with 12 substituted becomes

$$\sum_{s=1}^{n} x'^2_s = \sum_{s=1}^{n} \sum_{r=1}^{n} o_{sr} x_r x'_s = \sum_{r=1}^{n} x_r^2 \tag{14}$$

Reversing the order of summation in the double sum yields

$$\sum_{r=1}^{n} \left(\sum_{s=1}^{n} o_{sr} x'_s \right) x_r = \sum_{r=1}^{n} x_r x_r \tag{15}$$

from which it becomes clear that the stipulation expressed by Eq. 13 leads to the result

$$\sum_{s=1}^{n} o_{sr} x'_s = x_r \tag{16}$$

Written out for $r = 1, 2, \cdots n$, this reads

$$\left.\begin{array}{l} o_{11}x'_1 + o_{21}x'_2 + \cdots + o_{n1}x'_n = x_1 \\ o_{12}x'_1 + o_{22}x'_2 + \cdots + o_{n2}x'_n = x_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ o_{1n}x'_1 + o_{2n}x'_2 + \cdots + o_{nn}x'_n = x_n \end{array}\right\} \tag{17}$$

This is the inverse transformation with respect to the one expressed by Eqs. 12. Its matrix is seen to be the transpose of $\mathbf{O}$. A matrix whose transpose is equal to its inverse is shown in Art. 6, Ch. II, to have the properties expressed by Eq. 78 or more specifically by Eqs. 79 and 80 of that chapter.

The direction cosines of the new set of rectangular co-ordinate axes with respect to the old ones, therefore, satisfy the relations

$$o_{s1}o_{r1} + o_{s2}o_{r2} + \cdots + o_{sn}o_{rn} = \begin{cases} 1 & \text{for } s = r \\ 0 & \text{for } s \neq r \end{cases} \tag{18}$$

and

$$o_{1s}o_{1r} + o_{2s}o_{2r} + \cdots + o_{ns}o_{nr} = \begin{cases} 1 & \text{for } s = r \\ 0 & \text{for } s \neq r \end{cases} \tag{19}$$

In terms of the vector set of $\mathcal{O}$, that is, $o_1, o_2, \cdots o_n$ and the transposed vector set $o_1{}^t, o_2{}^t, \cdots o_n{}^t$, these relations may be expressed by the scalar products

$$o_s \cdot o_r = \begin{cases} 1 & \text{for } s = r \\ 0 & \text{for } s \neq r \end{cases} \qquad [20]$$

and

$$o_s{}^t \cdot o_r{}^t = \begin{cases} 1 & \text{for } s = r \\ 0 & \text{for } s \neq r \end{cases} \qquad [21]$$

It thus becomes clear that the vector set of the matrix belonging to the transformation from one rectangular co-ordinate system to another consists of a cluster of *mutually orthogonal* vectors of unit length. The same is true of the transposed vector set, which also is the vector set of the inverse matrix. The reason for designating this kind of matrix as an *orthogonal* one is thus evident. The vectors likewise are said to form an orthogonal set.

The relation 9, which is shown in the previous article to hold for the elements of any nonsingular matrix and those of its inverse, becomes identical with Eq. 21 when written for the elements of the orthogonal matrix and those of its inverse or transpose.

As pointed out in the discussion of the orthogonal matrix, the product

$$\mathcal{O} \times \mathcal{O}_t = \mathcal{O} \times \mathcal{O}^{-1} = \mathcal{U} \qquad [22]$$

Since the determinant of the unit matrix $\mathcal{U}$ is unity, and the determinant of $\mathcal{O}_t$ is the same as that of $\mathcal{O}$, it follows that

$$O^2 = 1, \qquad \text{or} \qquad O = \pm 1 \qquad [23]$$

A closer examination of the geometry involved shows that the value of the determinant O is $+1$ if the directional sequence of the vectors $o_1, o_2, \cdots o_n$ corresponds to a right-hand screw rule, and it is -1 if this sequence follows a left-hand screw rule. If both the reference axes 1, 2, $\cdots n$, and $1', 2', \cdots n'$ form right-hand or both left-hand systems, $O = +1$; but if one is a right- and the other a left-hand system, $O = -1$.

Besides being the matrix of a transformation representing the rotation of a rectangular system of reference axes, the orthogonal matrix is evidently also the matrix of a transformation for which a vector x and its transform y have the same length. In other words, a linear transformation may represent either a change from one co-ordinate system to another or the transformation from one point in space to another with respect to the same reference system. If the transformation is an orthogonal one, it represents a pure rotation about the origin; that is, a rotation either of the co-ordinate system for a fixed point, or of an arbitrary point in a fixed

system. In either case, the axes of the reference system are at right angles to each other if the transformation is an orthogonal one.

Such a rectangular co-ordinate system is commonly called a Cartesian system, though Descartes (after whom it is named) also devoted many of his studies to systems of co-ordinates in which the axes make oblique angles with each other. In a three-dimensional oblique reference system, the projection (co-ordinate) of a point on a given axis is found by dropping a line to this axis parallel to the plane of the other two axes (in a rectangular reference system this line is the perpendicular dropped from the point in question). An analogous interpretation applies to oblique systems of more dimensions.

Any linear nonsingular transformation may be regarded as a transformation of the co-ordinates of a point in a given oblique co-ordinate system to those of the same point in another system with coincident origin, or it may be regarded as the transformation from one point to another in the same oblique co-ordinate system.

An oblique co-ordinate system is also called an *affine* system. The transformation from one such system to another or from one point to another in the same system is sometimes referred to as a *linear affine transformation*. The orthogonal transformation is a special case of this more general type.

5. Transformation to an oblique co-ordinate system

In the transformation from one orthogonal co-ordinate system to another, the direction cosines of the axes $1'$ with respect to the axes $1, 2, \cdots n$ (these are the elements of the first row of the matrix $\mathfrak{O}$, Eq. 11) may be regarded as the components (projections) of a unit vector lying in the axis $1'$. This is the vector o_1 of the vector set of $\mathfrak{O}$. Similarly the vector o_2 of this set is seen to be a unit vector emanating from the common origin of the two orthogonal co-ordinate systems, coincident in direction with axis $2'$. The vector set $o_1, o_2, \cdots o_n$ of $\mathfrak{O}$ is thus seen to be a set of coterminus unit vectors coincident respectively with the orthogonal axes of the system $1', 2', \cdots n'$. By a similar line of reasoning, the transposed vector set $o_1{}^t, o_2{}^t, \cdots o_n{}^t$ is recognized as a set of coterminus unit vectors coincident respectively with the orthogonal axes of the co-ordinate system $1, 2, \cdots n$.

This helpful visualization or geometrical interpretation of the coefficients of a transformation matrix may be extended to the nonorthogonal matrix associated with the transformation to an oblique (affine) co-ordinate system, but several novel features enter into the geometrical interpretations, requiring further detailed discussion.

From the algebraic point of view, the number of dimensions of the

space under consideration is arbitrary, but for facilitating the geometrical visualization of the argument, reference is made to the two-dimensional Fig. 1. Here there are drawn three co-ordinate systems — one rectangular Cartesian system and two oblique systems. The axes of the rectangular system are designated by the letters X_1 and X_2. The first oblique system has the axes Ξ_1 and Ξ_2, and the second oblique system has the axes Ξ^*_1

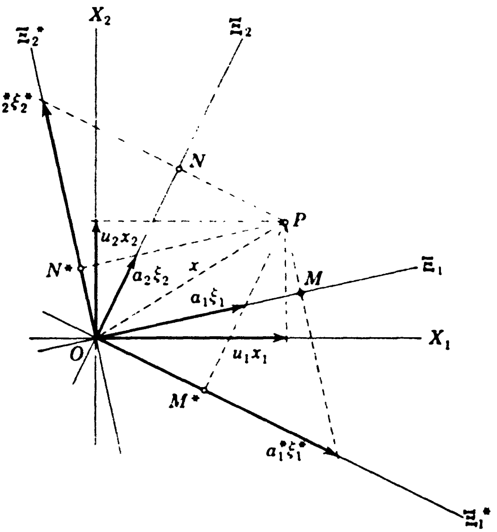


FIG. 1. The vector x described in equivalent fashion in oblique and rectangular co-ordinate systems.

and Ξ^*_2 which are at right angles to the axes Ξ_2 and Ξ_1 respectively of the first oblique system. The reason for considering the two oblique systems Ξ and Ξ^* (with the relative orientations of axes as just described) in conjunction with the rectangular system X follows from the discussion below.

For the rectangular co-ordinate system X, a set of unit vectors u_1, $u_2, \cdots u_n$ is defined which emanate from the common origin O and are coincident respectively with the axes $1, 2, \cdots n$ of this system, and hence are mutually at right angles with each other. A vector x from O to P has projections on the axes $1, 2, \cdots n$ which are denoted by $x_1, x_2, \cdots x_n$ respectively. The vector components of this vector are $u_1x_1, u_2x_2, \cdots u_nx_n$. The vector sum of these components, according to the usual parallelogram (or parallelopiped) rule of vector addition, yields the vector x, thus:

$$x = u_1x_1 + u_2x_2 + \cdots + u_nx_n \qquad [24]$$

In order to provide an equivalent representation for the vector x in the oblique system Ξ, a set of unit vectors $a_1, a_2, \cdots a_n$ is defined, which emanate from the common origin O and are coincident respectively with the axes $1, 2, \cdots n$ of this oblique system. The projections of the vector x (co-ordinates of the point P) upon these oblique axes are denoted by $\xi_1, \xi_2, \cdots \xi_n$ respectively. The vector components of x in the oblique system Ξ are, therefore, $a_1\xi_1, a_2\xi_2, \cdots a_n\xi_n$, and their vector sum also yields the vector x, that is,

$$x = a_1\xi_1 + a_2\xi_2 + \cdots + a_n\xi_n \tag{25}$$

Finally, the vector x may also be represented in the second oblique system Ξ^* by defining for it the set of unit vectors $a^*_1, a^*_2, \cdots a^*_n$ emanating from the common origin O and coinciding respectively with the axes $1, 2, \cdots n$ of this system. The projections of x upon these axes are denoted respectively by $\xi^*_1, \xi^*_2, \cdots \xi^*_n$. The vector components of x in the system Ξ^* are $a^*_1\xi^*_1, a^*_2\xi^*_2, \cdots a^*_n\xi^*_n$, and their vector sum yields

$$x = a^*_1\xi^*_1 + a^*_2\xi^*_2 + \cdots + a^*_n\xi^*_n \tag{26}$$

The unit vectors $u_1, u_2, \cdots u_n$ are regarded as the vector set of a matrix

$$\mathfrak{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{bmatrix} \tag{27}$$

in which the elements are the components of the vectors with respect to the rectangular co-ordinate system X. According to the definition of this vector set, it follows that this matrix is the unit matrix

$$\mathfrak{U} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & 1 \end{bmatrix} \tag{28}$$

The unit vectors $a_1, a_2, \cdots a_n$ of the co-ordinate system Ξ are regarded as the vector set of a matrix

$$\mathfrak{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{29}$$

in which the elements are the components of the vectors with respect to the rectangular system X also. For example, $a_{11}, a_{12}, \cdots a_{1n}$ are the projections of the vector a_1 upon the axes $1, 2, \cdots n$ respectively of the

system X; a_{21}, a_{22}, $\cdots$ a_{2n} are the projections of the vector a_2 upon the axes 1, 2, $\cdots$ n respectively of the system X, and so forth.

Similarly the unit vectors a^*_1, a^*_2, $\cdots$ a^*_n of the co-ordinate system Ξ^* are regarded as the vector set of a matrix

$$\mathfrak{A}^* = \begin{bmatrix} a^*_{11} & a^*_{12} & \cdots & a^*_{1n} \\ a^*_{21} & a^*_{22} & \cdots & a^*_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a^*_{n1} & a^*_{n2} & \cdots & a^*_{nn} \end{bmatrix} \qquad [30]$$

in which the elements again are the components of the vectors with respect to the rectangular system X. That is, a^*_{11}, a^*_{12}, $\cdots$ a^*_{1n} are the projections of the vector a^*_1 upon the axes 1, 2, $\cdots$ n respectively of the system X; a^*_{21}, a^*_{22}, $\cdots$ a^*_{2n} are the projections of the vector a^*_2 upon the axes 1, 2, $\cdots$ n respectively of the system X, and so forth.

According to the definition of the systems Ξ and Ξ^*, the axis 1 of Ξ is orthogonal to the axes 2, 3, $\cdots$ n of Ξ^*; the axis 2 of Ξ is orthogonal to the axes 1, 3, $\cdots$ n of Ξ^*, and so forth. Hence the scalar products of a_1 with the vectors a^*_2, a^*_3, $\cdots$ a^*_n are zero; the scalar products of the vector a_2 with the vector a^*_1, a^*_3, $\cdots$ a^*_n are zero, etc.

At this point a peculiarity of affine geometry (usually confusing to engineers) must be clearly understood. A distinction must be made between the scale by means of which any length is measured and the scales which are attached to the oblique co-ordinate axes. The "scale of length" is that which is attached to the rectangular Cartesian axes X (this is the same for all the axes 1, 2, $\cdots$ n in this system). The oblique axes carry their own scales but they are not used to measure length. The system Ξ may in general have a different scale for each of its axes, and they are different from the scales which are used to lay off the units on the axes of the system Ξ^*. The unit vectors a_1, a_2, $\cdots$ a_n are units according to the scales for the axes of the Ξ system, and the unit vectors a^*_1, a^*_2, $\cdots$ a^*_n are units according to the scales for the axes of the Ξ^* system, *but none of these unit vectors in general has unit length* according to the "scale of length," which is the scale for the units laid off on the axes of the system X.

The reader should appreciate that a scale on a co-ordinate axis may, in general, have a twofold purpose:

(a) It may be used to determine the *value* of the projection of a point upon that axis.

(b) It may be used as a tape measure is used to determine the distance between two points in space.

A scale may be used for either purpose alone, or for both. Use (a) is the one with which the reader is undoubtedly most familiar in connection with his analytic work. For example, when plotting a function $y = f(x)$,

scales are employed for the x- and y-axes which seldom have the same size of unit, and are used merely to read values of x and y, not also to measure length or distance.

Interest in the measurement of length arises only when geometrical considerations enter into the problem. When they do, one must have a "scale of length" (tape measure) in addition to the various scales whose use is restricted to the determination of the values of projections. In the present problem the identical scales carried by the axes of the system X serve also for the measurement of length. This designation having been made, it is obvious that no other scales (except if they are identical with those of the X-system) may be used for the measurement of length.

For curvilinear co-ordinates the situation is even more confusing, since the scales which the co-ordinate axes bear not only are in general different for the different axes of the same system, but also vary from point to point along these axes, whereas the scale of length is independent of the co-ordinate system and independent of the location within any system. As long as the oblique axes are linear, their scales are the same for all points within the system.

The units for the scales carried by the axes of the oblique systems may have their lengths (as measured by the scale of the X-system) so adjusted that the scalar products of the unit vectors $a_1, a_2, \cdots a_n$ *respectively* with the unit vectors $a^*_1, a^*_2, \cdots a^*_n$ are *all unity*. (This adjustment is called "normalization.") For example, if the angle between a_1 and a^*_1 is α_1, the lengths of a_1 and a^*_1 (that is, $|a_1|$ and $|a^*_1|$) are determined so that $|a_1| \times |a^*_1| \cos \alpha_1$ equals unity. Since α_1 is an arbitrary angle, it is clear that although the length of *either* a_1 *or* a^*_1 can be chosen equal to unity, both unit vectors certainly cannot have unit length.

As a result of this normalization process and the orthogonality between unlike numbered axes in the systems Ξ and Ξ^*, the sets of unit vectors in these systems satisfy the following conditions:

$$a_i \cdot a^*_k = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \qquad [31]$$

in which the scalar product is indicated by a dot. In terms of the matrices 29 and 30 these conditions are expressed by the matrix equation

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \times \begin{bmatrix} a^*_{11} & a^*_{21} & \vdots & a^*_{n1} \\ a^*_{12} & a^*_{22} & \vdots & a^*_{n2} \\ \cdot & \cdot & \vdots & \cdot \\ \cdot & \cdot & \vdots & \cdot \\ \cdot & \cdot & \vdots & \cdot \\ a^*_{1n} & a^*_{2n} & \vdots & a^*_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \cdots & 0 \\ 0 & 1 & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad [32]$$

which is equivalent to

$$\mathcal{A} \times \mathcal{A}^*_t = \mathcal{U} \qquad [33]$$

Hence it follows that

$$\mathcal{A}^*_t = \mathcal{A}^{-1} \qquad \text{or} \qquad \mathcal{A}^* = \mathcal{A}_t^{-1} \qquad [34]$$

The matrix $\mathcal{A}^*$ is, therefore, the reciprocal of $\mathcal{A}$. The vector set a^*_1, $a^*_2, \cdots a^*_n$ is called the reciprocal of the set $a_1, a_2, \cdots a_n$, and the oblique co-ordinate system Ξ^* is referred to as the reciprocal of the oblique system Ξ. The present result yields a useful geometrical interpretation for the elements of a square matrix and those of its reciprocal. Incidentally, it may be recalled in this connection that an orthogonal matrix is its own reciprocal, and this conclusion checks with the fact that an orthogonal co-ordinate system is also its own reciprocal.

By use of one of the three equivalent relations for the vector x given by Eqs. 24, 25, and 26, the length of this vector may be expressed in terms of the co-ordinates of the rectangular system X, or of either of the oblique systems Ξ or Ξ^*. The square of this length is evidently given by the scalar product of the vector x with itself. Using Eq. 24, one has

$$|x|^2 = x \cdot x = (u_1x_1 + u_2x_2 + \cdots + u_nx_n) \cdot (u_1x_1 + u_2x_2 + \cdots + u_nx_n) \qquad [35]$$

Because of the mutual orthogonality of the unit vectors, the scalar product of any vector with any other in the set is zero, whereas this product of any vector with itself is unity. Hence in terms of the co-ordinates of the rectangular system X,

$$|x|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \qquad [36]$$

which is the familiar Pythagorean proposition extended to n-dimensional space.

This result may be expressed in terms of the co-ordinates $\xi_1, \xi_2, \cdots \xi_n$ of the vector x in the oblique system Ξ by first determining the expressions for $x_1, x_2, \cdots x_n$ in terms of the ξ's, and then substituting these into Eq. 36. A general expression for the components $x_1, x_2, \cdots x_n$ is evidently given by

$$x_k = u_k \cdot x \quad \text{for} \quad k = 1, 2, \cdots n \qquad [37]$$

which states that the projections of x on the axes $1, 2, \cdots n$ of the system X are the scalar products of x with the unit vectors $u_1, u_2, \cdots u_n$ of this system.

Substituting Eq. 25 into Eq. 37 yields

$$x_k = u_k \cdot a_1\xi_1 + u_k \cdot a_2\xi_2 + \cdots + u_k \cdot a_n\xi_n \qquad [38]$$

But

$$u_k \cdot a_i = a_{ik} \tag{39}$$

because the scalar product of u_k with a_i equals the projection of the vector a_i upon axis k of the rectangular system X. Hence Eq. 38, written out for $k = 1, 2, \cdots n$, reads

$$\begin{aligned} x_1 &= a_{11}\xi_1 + a_{21}\xi_2 + \cdots + a_{n1}\xi_n \\ x_2 &= a_{12}\xi_1 + a_{22}\xi_2 + \cdots + a_{n2}\xi_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n &= a_{1n}\xi_1 + a_{2n}\xi_2 + \cdots + a_{nn}\xi_n \end{aligned} \tag{40}$$

This set of equations may be written in matrix form by defining the column matrices

$$x] = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{41}$$

and

$$\xi] = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \cdot \\ \cdot \\ \cdot \\ \xi_n \end{bmatrix} \tag{42}$$

With the transpose of the matrix 29, Eqs. 40 evidently are expressed by

$$x] = \mathcal{A}_t\, \xi] \tag{43}$$

Substitution of this result into Eq. 36 is now greatly facilitated by noting that

$$|x|^2 = x]_t \times x] = [x_1 x_2 \cdots x_n] \times \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = x_1{}^2 + x_2{}^2 + \cdots + x_n{}^2 \tag{44}$$

Thus the square of the length of the vector x, expressed in terms of the co-ordinates of the oblique system Ξ, becomes

$$|x|^2 = \xi]_t \times \mathcal{A}\mathcal{A}_t \times \xi] \tag{45}$$

Here it is effective to let

$$\mathcal{G} = \mathcal{A} \times \mathcal{A}_t = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix} \qquad [46]$$

The elements of this matrix are evidently expressed by the scalar products

$$g_{ik} = a_i \cdot a_k \qquad [47]$$

from which it is clear that the matrix $\mathcal{G}$ is symmetrical.

Equation 45 may now be written out, as follows:

$$\begin{aligned} |x|^2 = g_{11}\xi_1^2 &+ g_{12}\xi_1\xi_2 + \cdots + g_{1n}\xi_1\xi_n \\ + g_{21}\xi_2\xi_1 &+ g_{22}\xi_2^2 \;\; + \cdots + g_{2n}\xi_2\xi_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ + g_{n1}\xi_n\xi_1 &+ g_{n2}\xi_n\xi_2 + \cdots + g_{nn}\xi_n^2 \end{aligned} \qquad [48]$$

In the oblique system Ξ, the length of the vector x is *not* simply equal to the sum of the squares of its components in that system, but the result expressed by Eq. 48 shows that all cross-product terms are present as well. Because of the symmetry of the $\mathcal{G}$ matrix, which is expressed by

$$g_{ik} = g_{ki} \qquad [49]$$

it is clear that all the cross-product terms in Eq. 48 occur twice. This fact makes possible some condensation of the form of Eq. 48.

If all the scales for the axes of the Ξ system are chosen equal to the scale of length (as can be and usually is done), the unit vectors a_i have unit length. It then follows that all the elements on the principal diagonal of the matrix $\mathcal{G}$ become equal to unity. If the angle between the axes i and k in the Ξ system is denoted by θ_{ik}, the result in Eq. 47 shows that, for this choice of scales, the elements of $\mathcal{G}$ are given by the simple formula

$$g_{ik} = \cos\theta_{ik} \qquad [50]$$

which yields unity for $i = k$.

The determination of the matrix $\mathcal{G}$ is intimately connected with the choice of a scale in terms of which a length may be measured in the oblique co-ordinate system. *Length in this system is defined by the matrix* $\mathcal{G}$. The latter is the matrix of the so-called *fundamental metric tensor* whose components are the elements g_{ik}.

The present discussion should be recognized as fundamental to tensor algebra also, even though we are not concerned here with that closely related subject. Matrix equations and tensor equations, in fact, differ only in notation. The geometrical interpretations of them, as far as this

discussion is concerned, are identical. Tensor notation is indeed somewhat more advantageous than matrix notation for the operations now under consideration. The view taken in this book, however, is that matrix algebra is an invaluable aid to clear comprehension of the tensor method, and that the student will more readily assimilate that method once he has understood the essential ideas underlying the present discussion. Proper grasp of the tensor method comes not from the mere acquisition of a set of manipulative rules but rather from recognition that tensor algebra is a symbolic representation of geometrical and physical ideas.

The square of the length of the vector x given by Eq. 36 may also be expressed in terms of the co-ordinates ξ^* of the reciprocal oblique system Ξ^*. Thus, using Eq. 26, the co-ordinates of x in the rectangular system are, according to Eq. 37, given by

$$x_k = u_k \cdot a^*{}_1\xi^*{}_1 + u_k \cdot a^*{}_2\xi^*{}_2 + \cdots + u_k \cdot a^*{}_n\xi^*{}_n \qquad [51]$$

Here it is recognized that the scalar product

$$u_k \cdot a^*{}_i = a^*{}_{ik} \qquad [52]$$

so that Eq. 51 written out for $k = 1, 2, \cdots n$ reads

$$\begin{aligned} x_1 &= a^*{}_{11}\xi^*{}_1 + a^*{}_{21}\xi^*{}_2 + \cdots + a^*{}_{n1}\xi^*{}_n \\ x_2 &= a^*{}_{12}\xi^*{}_1 + a^*{}_{22}\xi^*{}_2 + \cdots + a^*{}_{n2}\xi^*{}_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n &= a^*{}_{1n}\xi^*{}_1 + a^*{}_{2n}\xi^*{}_2 + \cdots + a^*{}_{nn}\xi^*{}_n \end{aligned} \qquad [53]$$

With the definition of the column matrix

$$\xi^*] = \begin{bmatrix} \xi^*{}_1 \\ \xi^*{}_2 \\ \cdot \\ \cdot \\ \cdot \\ \xi^*{}_n \end{bmatrix} \qquad [54]$$

and the transpose of the reciprocal matrix 30, Eqs. 53 are given in matrix form by

$$x] = \mathfrak{A}^*{}_t\xi^*] \qquad [55]$$

Utilizing again the relation expressed by Eq. 44, the square of the length of the vector x in terms of the co-ordinates of the reciprocal oblique system reads

$$|x|^2 = \xi^*]_t \times \mathfrak{A}^*\mathfrak{A}^*{}_t \times \xi^*] \qquad [56]$$

Here it is effective to introduce the matrix

$$\mathcal{G}^* = \mathcal{A}^* \times \mathcal{A}^*_t = \begin{bmatrix} g^*_{11} & g^*_{12} & \cdots & g^*_{1n} \\ g^*_{21} & g^*_{22} & \cdots & g^*_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ g^*_{n1} & g^*_{n2} & \cdots & g^*_{nn} \end{bmatrix} \quad [57]$$

The elements of this matrix are given by the scalar products

$$g^*_{ik} = a^*_i \cdot a^*_k = g^*_{ki} \quad [58]$$

In view of Eqs. 34 and 46, it is recognized that

$$\mathcal{G}^* = \mathcal{A}_t^{-1} \times \mathcal{A}^{-1} = (\mathcal{A} \times \mathcal{A}_t)^{-1} = \mathcal{G}^{-1} \quad [59]$$

The matrix $\mathcal{G}^*$, which is fundamental to the measurement of length in the reciprocal oblique system, is the inverse of the matrix $\mathcal{G}$. Hence

$$g^*_{ik} = \frac{G_{ik}}{G} \quad [60]$$

in which G is the determinant of $\mathcal{G}$, and G_{ik} its cofactors. Alternatively,

$$g_{ik} = \frac{G^*_{ik}}{G^*} \quad [61]$$

in which G^* is the determinant of $\mathcal{G}^*$ and G^*_{ik} its cofactors. Evidently, by Eqs. 46 and 57,

$$G = A^2 = \frac{1}{G^*} \quad [62]$$

A representation for the elements g^*_{ik} similar in form to that given by Eq. 50 for g_{ik} is, according to Eq. 58,

$$g^*_{ik} = |a^*_i| \times |a^*_k| \cos \theta^*_{ik} \quad [63]$$

in which $|a^*_i|$ denotes the length of a unit vector in the reciprocal coordinate system Ξ^*, and θ^*_{ik} is the angle between the axes i and k in this system.

In general, the lengths* of the unit vectors in the oblique systems are given by the expressions

$$|a_i| = \sqrt{a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2} = \sqrt{g_{ii}} \quad [64]$$

and

$$|a^*_i| = \sqrt{a^*_{i1}{}^2 + a^*_{i2}{}^2 + \cdots + a^*_{in}{}^2} = \sqrt{g^*_{ii}} \quad [65]$$

*The term " magnitude " would actually be more appropriate here than " length " because the unit vectors a_i and a^*_i, like those for the rectangular coordinates, are assumed to be without dimensions.

When the former are chosen equal to unity, the latter are in general all different from unity; that is, each axis of the Ξ^* system carries a different scale.

The result given by Eq. 56 for the square of the length of the vector x in terms of the co-ordinates of the reciprocal oblique system, when written out, reads

$$\begin{aligned} |x|^2 = {} & g^*_{11}\xi^{*2}_1 + g^*_{12}\xi^*_1\xi^*_2 + \cdots + g^*_{1n}\xi^*_1\xi^*_n \\ & + g^*_{21}\xi^*_2\xi^*_1 + g^*_{22}\xi^{*2}_2 + \cdots + g^*_{2n}\xi^*_2\xi^*_n \\ & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ & + g^*_{n1}\xi^*_n\xi^*_1 + g^*_{n2}\xi^*_n\xi^*_2 + \cdots + g^*_{nn}\xi^{*2}_n \end{aligned} \qquad [66]$$

Again all cross-product terms (double products because $g_{ik} = g^*_{ki}$) are present in addition to the square terms.

Still another form for $|x|^2$ is possible if in Eq. 44 *both* of the alternative expressions 43 and 55 for $x]$ are substituted, one for $x]_t$ and the other for $x]$. Then the result reads

$$|x|^2 = \xi]_t \times \mathcal{A}\mathcal{A}^*_t \times \xi^*] \qquad [67]$$

or

$$|x|^2 = \xi^*] \times \mathcal{A}^*\mathcal{A}_t \times \xi] \qquad [68]$$

But by Eq. 33

$$\mathcal{A} \times \mathcal{A}^*_t = \mathcal{A}^* \times \mathcal{A}_t = \mathfrak{U} \qquad [69]$$

so that

$$\begin{aligned} |x|^2 &= \xi]_t \times \xi^*] = \xi^*]_t \times \xi] \\ &= \xi_1\xi^*_1 + \xi_2\xi^*_2 + \cdots + \xi_n\xi^*_n \end{aligned} \qquad [70]$$

In this expression, no cross-product terms are present. It closely resembles the simple Pythagorean form given by Eq. 36, the only difference being that products of corresponding components of the vector x in the two oblique systems appear in place of the squares of one kind of component.

The ξ_k are called the ***contravariant*** components and the ξ^*_k the ***covariant**** components of the vector x. The contravariant components are the components of x in the oblique system Ξ, and the covariant components are those in the reciprocal system Ξ^*. Note in this connection that, for the two oblique co-ordinate systems, the property of being

*These names are chosen with regard to the manner in which the components behave when subjected to a co-ordinate transformation. Thus the sets of variables $\xi_1 \cdots \xi_n$ and $\xi^*_1 \cdots \xi^*_n$ are said to be *contragredient* because when one set is subjected to a nonsingular linear transformation, the other is subjected to a linear transformation with the reciprocal matrix. For the moment, these matters need not be of interest, but they may be demonstrated from Eqs. 92 and 97 which follow.

reciprocal is strictly a *mutual* one. Hence the ξ^*_k's may just as well be looked upon as the contravariant components of x, in which case the ξ_k's become the covariant components.

According to Eqs. 43 and 55,

$$\mathcal{A}_t\xi] = \mathcal{A}^*{}_t\xi^*] \qquad [71]$$

Hence, by Eqs. 34 and 46,

$$\mathcal{A}\mathcal{A}_t\xi] = \mathcal{G}\xi] = \xi^*] \qquad [72]$$

It is thus seen that the covariant and contravariant components of the vector x are related by the matrix $\mathcal{G}$.

Another interesting pair of relationships is obtained through forming the scalar products $a_k \cdot x$ and $a^*_k \cdot x$, using for x the Eqs. 26 and 25 respectively, and noting the conditions 31. This procedure gives

$$a_k \cdot x = \xi^*_k \qquad [73]$$

and

$$a^*_k \cdot x = \xi_k \qquad [74]$$

With reference to Fig. 1, $a_k \cdot x$ for $k = 1$ and 2 are the orthogonal projections OM and ON of the vector x upon the axes Ξ_1 and Ξ_2. Equation 73 states that the lengths of these projections, measured with the scales of the Ξ_1 and Ξ_2 axes, are numerically equal to the covariant components ξ^*_1 and ξ^*_2 of x, and may be substituted for these in Eq. 70 in evaluating the length of the vector x. Similarly, $a^*_k \cdot x$ for $k = 1$ and 2 are the orthogonal projections OM^* and ON^* of the vector x upon the axes Ξ^*_1 and Ξ^*_2. Equation 74 states that the lengths of these projections, measured with the scales of the Ξ^*_1 and Ξ^*_2 axes, are numerically equal to the contravariant components ξ_1 and ξ_2 of x.

In view of the relation 73, the covariant components of the vector x are sometimes designated as the *orthogonal* projections of the point P in Fig. 1 upon the axes 1 and 2 of the system Ξ (as contrasted with the *parallel* projections which are the contravariant components). Although this procedure is justified numerically, and does away with the necessity of considering the reciprocal axes (or drawing them in the case of a graphical determination), it fails to give the true geometrical picture regarding the nature of the covariant components.

The scalar product of two vectors x and y in terms of their co-ordinates in an oblique system is readily determined with the aid of the above considerations. The components of y in the systems X, Ξ, and Ξ^* may be denoted by y_k, η_k, and η^*_k. The corresponding column matrices which

represent these components are

$$y] = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad \eta] = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \cdot \\ \cdot \\ \cdot \\ \eta_n \end{bmatrix} \qquad \eta^*] = \begin{bmatrix} \eta^*_1 \\ \eta^*_2 \\ \cdot \\ \cdot \\ \cdot \\ \eta^*_n \end{bmatrix} \quad . \qquad [75]$$

The matrix Equations 43 and 55 written for x and y

$$x] = \mathfrak{A}_t\xi] = \mathfrak{A}^*_t\xi^*] \qquad [76]$$

$$y] = \mathfrak{A}_t\eta] = \mathfrak{A}^*_t\eta^*] \qquad [77]$$

admit of four alternative representations for the scalar product

$$x \cdot y = x]_t \times y] \qquad [78]$$

They are

$$x \cdot y = \xi]_t \times \mathfrak{A}\mathfrak{A}_t \times \eta] = \xi]_t \times \mathfrak{G} \times \eta] \qquad [79]$$

$$= \xi^*]_t \times \mathfrak{A}^*\mathfrak{A}^*_t \times \eta^*] = \xi^*]_t \times \mathfrak{G}^* \times \eta^*] \qquad [80]$$

$$= \xi]_t \times \mathfrak{A}\mathfrak{A}^*_t \times \eta^*] = \xi]_t\eta^*] \qquad [81]$$

$$= \xi^*]_t \times \mathfrak{A}^*\mathfrak{A}_t \times \eta] = \xi^*]_t\eta] \qquad [82]$$

The first two of these results are similar in form to the Eqs. 48 and 66 except that the terms are bilinear in ξ_i and η_k (respectively ξ^*_i and η^*_k) instead of being quadratic in ξ_i (respectively ξ^*_i). The last two are the mixed forms, which read

$$x \cdot y = \xi_1\eta^*_1 + \xi_2\eta^*_2 + \cdots + \xi_n\eta^*_n \qquad [83]$$

$$x \cdot y = \xi^*_1\eta_1 + \xi^*_2\eta_2 + \cdots + \xi^*_n\eta_n \qquad [84]$$

The first of these involves the contravariant components of x and the covariant components of y; in the second form these two types of components are interchanged.

Except for the appearance of the two kinds of components, the expressions 83 and 84 parallel the customary one in rectangular co-ordinates, namely

$$x \cdot y = x_1y_1 + x_2y_2 + \cdots + x_ny_n \qquad [85]$$

In a rectangular (orthogonal) co-ordinate system, the covariant and contravariant components are identical, since the system is its own reciprocal.

6. Transformation from one oblique system to another

In addition to the oblique co-ordinate system characterized in terms of the rectangular system X by means of the matrix $\mathfrak{A}$, a second oblique

system may be considered as characterized by a matrix $\mathcal{B}$ with the set of unit vectors

$$b_k = u_1 b_{k1} + u_2 b_{k2} + \cdots + u_n b_{kn} \qquad (k = 1, 2, \cdots n) \qquad [86]$$

The reciprocal of this second oblique system has the set of unit vectors

$$b^*_k = u_1 b^*_{k1} + u_2 b^*_{k2} + \cdots + u_n b^*_{kn} \qquad (k = 1, 2, \cdots n) \qquad [87]$$

which is the vector set of the reciprocal matrix $\mathcal{B}^*$, that is,

$$b_i \cdot b^*_k = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases} \qquad [88]$$

The contravariant and covariant components of the vector x in the second oblique system are denoted by ζ_k and ζ^*_k, with the column matrices $\zeta]$ and $\zeta^*]$. Then the vector x, in terms of its components in the rectangular system X, has the following representations:

$$\begin{aligned} x] &= \mathcal{A}_t \xi] = \mathcal{A}^*_t \xi^*] \\ &= \mathcal{B}_t \zeta] = \mathcal{B}^*_t \zeta^*] \end{aligned} \qquad [89]$$

These relations readily yield the transformations from either of the variables ξ_k or ξ^*_k to either of the variables ζ_k or ζ^*_k associated with the second oblique system and its reciprocal.

For example, from the matrix equation

$$\mathcal{A}_t \xi] = \mathcal{B}^*_t \zeta^*] \qquad [90]$$

premultiplication by

$$\mathcal{B} = (\mathcal{B}^*_t)^{-1} \qquad [91]$$

gives

$$\mathcal{B}\mathcal{A}_t \times \xi] = \zeta^*] \qquad [92]$$

which relates the contravariant components of x in the first oblique system (these are the ξ_k's) to the covariant components of x in the second oblique system. The resultant transformation matrix in this case is

$$\mathcal{C} = \mathcal{B} \times \mathcal{A}_t \qquad [93]$$

If this is written out as

$$\mathcal{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \qquad [94]$$

with the vector

$$c_k = u_1 c_{k1} + u_2 c_{k2} + \cdots + u_n c_{kn} \qquad (k = 1, 2, \cdots n) \qquad [95]$$

then it is clear from Eq. 93 that the elements of $\mathcal{C}$ are given by the scalar products

$$c_{ik} = b_i \cdot a_k \tag{96}$$

Hence the elements of the resultant transformation matrix in Eq. 92 are seen to be the orthogonal projections of the unit vectors a_k upon the axes of the second oblique system (measured with the scales belonging to the axes of this system), or alternatively they may be considered as the projections of the unit vectors b_k upon the axes of the first oblique system, measured with the corresponding scales carried by these axes.

The transformation 92, with its resultant matrix determined as just described, transforms the contravariant components of x in the first system into covariant components in the second oblique system.

Because of the mutual character of the relation between an oblique co-ordinate system and its inverse, as pointed out in the previous article, Eq. 92 evidently remains true if $\mathcal{B}$ is replaced by $\mathcal{B}^*$ and simultaneously ζ is replaced by ζ^*. This shift gives

$$\mathcal{B}^*\mathcal{A}_t \times \xi] = \zeta] \tag{97}$$

which is an equation relating the contravariant components of x in the first oblique system to the *same* kind of components in the second system. The elements of the resultant transformation matrix are determined and interpreted geometrically just as for the transformation 92 except that the *reciprocal* of the second oblique system with its unit vectors b^*_k replaces the unit vector b_k with its corresponding axes.

Further detailed discussion of these transformations becomes decidedly awkward in terms of matrix notation, as the reader can at this point readily appreciate by continuing the formation of various additional obvious relationships. Indeed, it is this circumstance which is the best justification for the introduction of the notation and the conventions of tensor algebra. The most important point in this notation lies in the designation of components in the reciprocal co-ordinate systems (covariant quantities) by *lower* indexes (subscripts) and those in the given co-ordinate systems (contravariant quantities) by *upper* indexes (superscripts) instead of by the placing of an asterisk on these quantities. The circumspection which results from this simple artifice alone may readily be appreciated even by the reader who has as yet no familiarity with the tensor notation.

However, a detailed discussion of these matters at this point would lead the reader too far afield from the present objectives, which are to lay a general foundation for a more thorough consideration of this subject at a later time.

7. Systems of linear algebraic equations

As a result of the geometrical interpretations given in the preceding articles the problem of finding solutions to a set of linear simultaneous algebraic equations may be discussed with a greater degree of clarity than is possible from a purely algebraic point of view. To start with, the number of equations is considered equal to the number of unknowns. The equations read

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned} \tag{98}$$

with the matrix

$$\mathcal{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{99}$$

The components a_{ik}, x_k, and y_k of the vector set a_i and the vectors x and y respectively for $k = 1, 2, \cdots n$ are the ordinary components with respect to a rectangular Cartesian co-ordinate system in an n-dimensional space.

The vector set a_i (for $i = 1, 2, \cdots n$) and the vector y are specified, and the problem is to find the corresponding vector x. The problem of particular interest according to the discussion in Art. 9, Ch. I, is the question whether a vector x can be found to satisfy the transformation 98 when y is specified to be zero. In other words: Can the transform of a vector x be zero?

In this case Eqs. 98 become homogeneous. In vector form they read

$$\begin{aligned} a_1 \cdot x &= 0 \\ a_2 \cdot x &= 0 \\ \cdots\cdots \\ a_n \cdot x &= 0 \end{aligned} \tag{100}$$

According to these equations, the vector x (if one exists) *must be simultaneously orthogonal to all the vectors* $a_1, a_2, \cdots a_n$ *in the vector set of* $\mathcal{A}$. The possibility of the existence of such a vector evidently depends upon the rank of this vector set. If the rank is n, then, according to the discussion in Art. 2 above, the vector set utilizes all the available n dimensions, and it is clear that no vector x can be found to satisfy the conditions 100. In this case the determinant of $\mathcal{A}$ is not zero.

On the other hand, if the rank of the vector set of $\mathcal{A}$ is $(n - 1)$, one of the available n dimensions is left unoccupied. For example, if in three-dimensional space the vectors a_1, a_2, and a_3 lie in a plane, only two dimen-

sions are utilized by this vector set, and any vector x normal to this plane constitutes a solution to the homogeneous Eqs. 100. The vanishing of the determinant of $\mathcal{A}$ is seen to be the necessary and sufficient condition for the existence of such a solution.

At the same time it is also clear that the solution merely specifies a direction for the vector x; its length is arbitrary. This checks with the conclusion reached in Art. 9, Ch. I, on purely algebraic grounds, according to which only the ratios of the unknowns $x_1, x_2, \cdots x_n$ to each other are determined from the homogeneous equations. The direction cosines of the vector x are given by the expression

$$\frac{x_k}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}} = \frac{x_k/\sqrt{x_k^2}}{\sqrt{\dfrac{x_1^2}{x_k^2} + \dfrac{x_2^2}{x_k^2} + \cdots + \dfrac{x_n^2}{x_k^2}}} \qquad [101]$$

for $k = 1, 2, \cdots n$, and this expression involves only the ratios of the unknowns.

If the rank of the vector set of $\mathcal{A}$ is $(n - 2)$, two of the available dimensions are left unoccupied. For example, if in three-dimensional space the vectors a_1, a_2, and a_3 lie in a straight line, any vector x lying in the plane normal to this line is simultaneously orthogonal to all three vectors in the set, and hence any such vector constitutes a solution. Although an infinite number of x-vectors may be drawn in this plane, only two independent ones can be found in this way. Hence when the rank of the matrix $\mathcal{A}$ is $(n - 2)$, *two* independent solutions exist to the set of homogeneous Eqs. 100.

In general, then, if the rank of $\mathcal{A}$ is $(n - p)$, the homogeneous equations have p independent solutions. In any case, however, the length of the vector x for any of the solutions remains arbitrary; and if the rank of $\mathcal{A}$ is $r < n$, the equations determine only r of the x_k's in terms of the remaining ones. Choosing arbitrary sets of values for these remaining x_k's permits any number of solutions to be obtained, of which no more than p can be linearly independent. The x-vectors corresponding to p independent solutions are in general not mutually orthogonal to each other, but a set of p mutually orthogonal solutions may be found by proceeding in the following manner.

If the vector set $a_1, a_2, \cdots a_n$ has the rank r, it is surely possible to find r linearly independent vectors in this set. The numbering in the original set may be changed if necessary so that the vectors $a_1, a_2, \cdots a_r$ are the linearly independent ones. Then the r equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rn}x_n &= 0 \end{aligned} \qquad [102]$$

uniquely determine r of the x_k's in terms of the remaining ones. That is, a set of values

$$x_k = \alpha_{1k} \qquad (\text{for } k = 1, 2, \cdots n) \tag{103}$$

of which $p = n - r$ are arbitrarily chosen (not more than $p - 1$ of these may be chosen equal to zero, however) may be considered to be the components of the first x-vector constituting a solution. This is the vector $x = \alpha_1$ with the components α_{1k}.

Now the vector set $a_1 \cdots a_r$ is augmented by having the vector α_1 associated with it. This augmented set is surely an independent one, because α_1 is orthogonal to all the vectors $a_1 \cdots a_r$. Hence the $r + 1$ equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rn}x_n &= 0 \\ \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= 0 \end{aligned} \tag{104}$$

uniquely determine $r + 1$ of the x_k's in terms of the remaining ones. That is, a set of values

$$x_k = \alpha_{2k} \qquad (\text{for } k = 1, 2, \cdots n) \tag{105}$$

of which $p - 1$ are arbitrarily chosen (not more than $p - 2$ of these may be chosen equal to zero) represents the components of a vector x constituting a second solution which is orthogonal to the augmented vector set $a_1 \cdots a_r, \alpha_1$, and hence orthogonal to the first solution $x = \alpha_1$. This second solution is the vector $x = \alpha_2$ with the components α_{2k}.

A third vector $x = \alpha_3$ with the components α_{3k} is found in like manner by solving the set of $r + 2$ homogeneous equations corresponding to the linearly independent vector set $a_1 \cdots a_r, \alpha_1, \alpha_2$. These solutions are the components

$$x_k = \alpha_{3k} \qquad (\text{for } k = 1, 2, \cdots n) \tag{106}$$

of which $p - 2$ are arbitrarily chosen.

A continuation of this process finally yields a set of mutually orthogonal vectors $\alpha_1, \alpha_2, \cdots \alpha_p$, all of which are orthogonal to all the vectors in the set $a_1, a_2, \cdots a_r$. Such a set of mutually orthogonal solutions is, however, not unique, because of the arbitrary choice of p, $p - 1$, $p - 2$, etc., x_k-values at the various steps in the process of solution as described above.

The case for which the rank of the vector set of $\mathcal{A}$ is $r = n - 1$ is most frequent in its occurrence in practice. Here only one independent solution (direction for the x-vector) exists. Because of the relation expressed by Eq. 46 of Ch. I and because of the fact that $A = 0$ for $r = n - 1$, the

following relationship holds in this case (for *all* values of i and k from 1 to n)

$$a_{i1}A_{k1} + a_{i2}A_{k2} + \cdots + a_{in}A_{kn} = 0 \qquad [107]$$

in which A_{ks} are the cofactors of the elements a_{ks} formed from the determinant A of the matrix $\mathfrak{A}$.

If A_k is defined as a vector having the components A_{ks}, the relation 107 in vector form reads

$$a_i \cdot A_k = 0 \qquad [108]$$

which states that the vector A_k is simultaneously orthogonal to all the vectors in the set $a_1, a_2, \cdots a_n$. This vector, therefore, constitutes the desired solution.

Note that the cofactors A_{ks} for $s = 1, \cdots n$ are formed for the elements of the kth row of $\mathfrak{A}$. Any row may be chosen. If, however, all the cofactors of the elements of that row happen to be zero (as is possible because the rank is $n - 1$), the solution assumes an indeterminate form. However, since the rank is assumed for this case to be not less than $n - 1$, a row can surely be found for which not all the cofactors of the elements are zero.

The direction cosines of the vector x constituting the desired solution are given by

$$\cos\theta_s = \frac{A_{ks}}{\sqrt{A_{k1}^2 + A_{k2}^2 + \cdots + A_{kn}^2}} \qquad [109]$$

in which θ_s is the angle between the direction of the vector x and the axis s of the rectangular co-ordinate system.

Another problem of special interest in connection with Eqs. 98 occurs when the rank of the vector set of $\mathfrak{A}$ is less than n, but the y_i-values on the right-hand sides of these equations do not happen to be zero. That is, the equations are inhomogeneous but the determinant of their matrix is zero. In Art. 9, Ch. I, it is pointed out that solutions may still exist if the y_i's have values satisfying Eq. 56, which reads

$$y_i = \alpha_1 a_{i1} + \alpha_2 a_{i2} + \cdots + \alpha_n a_{in} \qquad (i = 1, 2, \cdots n) \qquad [110]$$

but that the solutions assume an indeterminate form when expressed according to Cramer's rule.

The α_k's in Eq. 110 are arbitrary factors. In vector form this equation reads

$$y = \alpha_1 a_1^t + \alpha_2 a_2^t + \cdots + \alpha_n a_n^t \qquad [111]$$

and states that the vector y is given by a linear combination of the transposed vector set of $\mathfrak{A}$. By making use of the relations 110, Eqs. 98

may be rewritten in the form

$$
\begin{aligned}
a_{11}x'_1 + a_{12}x'_2 + \cdots + a_{1n}x'_n &= 0 \\
a_{21}x'_1 + a_{22}x'_2 + \cdots + a_{2n}x'_n &= 0 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{n1}x'_1 + a_{n2}x'_2 + \cdots + a_{nn}x'_n &= 0
\end{aligned}
\qquad [112]
$$

in which

$$x'_k = x_k \quad \alpha_k \qquad [113]$$

Or if $\alpha_1, \alpha_2, \cdots \alpha_n$ are considered to be the components of a vector α, and $x'_1, x'_2, \cdots x'_n$ the components of a vector x',

$$x' = x - \alpha \qquad \text{and} \qquad x = x' + \alpha \qquad [114]$$

The vector x' is a solution to the homogeneous equations 112. The problem of finding solutions x' when the rank of $\mathcal{A}$ is $r < n$ is discussed above, and the desired values of the vector x are given in terms of those for x' by Eq. 114.

When the vector y, in this type of problem (for which $r < n$), cannot be expressed in the form of Eq. 111 (in this case y is not a member* of the transposed vector set of $\mathcal{A}$), the original set, Eqs. 98, is said to be *inconsistent*, and no solutions are possible.

It is interesting that the discussion of this special problem incidentally offers also an alternative interpretation to the solution of the general inhomogeneous equations 98. This case is contained in the solutions to the homogeneous equations 112 for x' when the rank of $\mathcal{A}$ is n. Then the application of Cramer's rule yields $x' = 0$; that is, $x'_1 = x'_2 = \cdots x'_n = 0$, and the desired solution for x follows from Eq. 114, namely, for $r = n$,

$$x = \alpha \qquad [115]$$

In this case, Eq. 111 yields the following representation for the vector y:

$$y = x_1a_1{}^t + x_2a_2{}^t + \cdots + x_na_n{}^t \qquad [116]$$

which is an interesting alternative form for the linear nonsingular transformation.

It appears from this discussion that the necessary and sufficient condition for the existence of solutions to the inhomogeneous equations 98 is that the vector y be a member of the transposed vector set $a_1{}^t, a_2{}^t, \cdots a_n{}^t$, regardless of the rank of $\mathcal{A}$. If the rank is n, then y must necessarily be a member of the transposed vector set (no matter what the values of $y_1 \cdots y_n$ may be) because the rank of the augmented set $a_1{}^t, a_2{}^t, \cdots a_n{}^t, y$ cannot be greater than n (see footnote below). Then

*If the rank of the transposed vector set and of that set augmented by the vector y are alike, then y is called a *member* of the transposed vector set, and hence is expressible in the form given by Eq. 111.

there exists a unique solution for x which is given by Cramer's rule. When the rank of $\mathcal{A}$ is less than n, no solutions exist unless y is a member of the transposed vector set of $\mathcal{A}$, in which case $n - r$ linearly independent solutions can be found.

A summary statement regarding the possibility of finding solutions to Eqs. 98, known as *the rule of alternatives*, may now be formulated, as follows:

Either the matrix $\mathcal{A}$ of Eqs. 98 has the rank n, whence $A \neq 0$, and there exists one set of values $x_1, x_2, \cdots x_n$ which satisfies the equations for arbitrary values of $y_1, y_2, \cdots y_n$—in particular the trivial values $x_1 = x_2 = \cdots = x_n = 0$ resulting for $y_1 = y_2 = \cdots = y_n = 0$ – these solutions being given by Cramer's rule,

or the matrix $\mathcal{A}$ has the rank $r = n - p$ $(p \neq 0)$, whence $A = 0$, and the corresponding homogeneous equations have p independent solutions. The inhomogeneous equations then have solutions for particular values of $y_1, y_2, \cdots y_n$ only, namely those values for which the vector y is expressible as a linear combination of the transposed vector set of $\mathcal{A}$. To these particular solutions may be added those for the corresponding homogeneous system.

The more general case of m equations with n unknowns

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = y_m \end{array} \qquad [117]$$

may now readily be disposed of. Multiplying the equations respectively by the unit vectors $u_1, u_2, \cdots u_m$ associated with the m axes of a rectangular co-ordinate system, and adding, yields the result

$$y = x_1 a_1{}^t + x_2 a_2{}^t + \cdots + x_n a_n{}^t \qquad [118]$$

where

$$y = u_1 y_1 + u_2 y_2 + \cdots + u_m y_m \qquad [119]$$

and

$$a_k{}^t = u_1 a_{1k} + u_2 a_{2k} + \cdots + u_m a_{mk} \qquad [120]$$

are the vectors in the transposed set of the matrix $\mathcal{A}$ of Eqs. 117.

The result given by Eq. 118 states that for any set of values $x_1, x_2, \cdots x_n$ which simultaneously satisfies the m equations 117, the vector y is a member of the transposed vector set of $\mathcal{A}$. In other words, unless y is a member of this vector set no solutions are possible, regardless of the rank of the set. This condition, which is arrived at above for the special

case $m = n$, is thus seen also to be the condition for the existence of solutions in the more general case $m \neq n$.

When $m \neq n$ this condition is more significant, for if the rank is n and $m = n$, the condition 118 is automatically fulfilled; whereas when $m > n$, for example, and the rank is n, y may or may not be a member of the transposed vector set. In other words, when $m = n$ and the rank is n, there is no question about the existence of solutions to the inhomogeneous equations; but when $m \neq n$ and the rank is $n(<m)$ or $m(<n)$, the existence of solutions still depends upon the fulfillment of the condition 118.

If this condition is written

$$y = \alpha_1 a_1{}^t + \alpha_2 a_2{}^t + \cdots + \alpha_n a_n{}^t \tag{121}$$

and $\alpha_1, \alpha_2, \cdots \alpha_n$ are regarded as the components of a vector α, then by letting

$$x' = x - \alpha \qquad x = x' + \alpha \tag{122}$$

as in the problem for $m = n$, the question of solutions to Eqs. 117 may be discussed for the inhomogeneous and the homogeneous cases simultaneously, because the resulting equations

$$\begin{array}{l} a_{11}x'_1 + a_{12}x'_2 + \cdots + a_{1n}x'_n = 0 \\ a_{21}x'_1 + a_{22}x'_2 + \cdots + a_{2n}x'_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x'_1 + a_{m2}x'_2 + \cdots + a_{mn}x'_n = 0 \end{array} \tag{123}$$

apply to the inhomogeneous equations 117 for $\alpha \neq 0$ and to the corresponding homogeneous equations for $\alpha = 0$.

As discussed in Art. 2 of this chapter, the rank of the nonsquare matrix $\mathfrak{A}$ of this set of equations is at most equal to the smaller of the two numbers m and n. If $m > n$, then $r \leqq n$. Out of the m equations 123 it is possible to select r independent ones (the coefficients of these define r independent vectors in the vector set of $\mathfrak{A}$). For these equations, $n - r$ independent solutions x' may be found by the procedure discussed for the case $m = n$ and $r \leqq n$. In particular, for $r = n$ only the trivial solution $x' = 0$ exists, which yields $x = \alpha$. The solutions so found automatically satisfy the remaining $m - r$ equations in the set 123 because the coefficients of these equations are the components of those vectors in the vector set of $\mathfrak{A}$ which are linear combinations of the others.

When $m < n$, then $r \leqq m$. Equations 123 then yield $n - r$ independent solutions which again are found as for the case $m = n$ and $r \leqq m$. In particular, for $r = m$ there are $n - m$ independent solutions.

8. On the rank of matrices having a null product

If $\mathcal{A}$ and $\mathcal{B}$ are two square matrices of order n, their product

$$\mathcal{C} = \mathcal{A} \times \mathcal{B} \tag{124}$$

is also a square matrix of order n with coefficients given by the scalar products

$$c_{ik} = a_i \cdot b_k^t \tag{125}$$

in which $a_1, a_2, \cdots a_n$ is the vector set of $\mathcal{A}$ and $b_1^t, b_2^t, \cdots b_n^t$ is the transposed vector set of $\mathcal{B}$.

All the vectors a_i can be orthogonal to the vectors b_k^t, in which case all the coefficients c_{ik} are zero. It thus becomes clear that the matrix product $\mathcal{A} \times \mathcal{B}$ may be zero even though neither $\mathcal{A}$ nor $\mathcal{B}$ is zero. This is an important difference between the laws of matrix algebra and those of scalar algebra, according to which the relation $ab = 0$ requires that either a or b be zero.

The vector interpretation of Eq. 125 allows the following conclusions to be drawn when $\mathcal{C} = 0$:

(a) If $\mathcal{A}$ is of rank n, then $\mathcal{B}$ must have the rank zero; that is, in this case $\mathcal{B} = 0$.
(b) If neither $\mathcal{A}$ nor $\mathcal{B}$ is zero, then the rank of both must be less than n.

The first of these conclusions follows from the consideration that if the vector set a_i has the rank n (occupies all the available dimensions) no vector exists which is simultaneously orthogonal to all the vectors a_i; that is, $b_k^t = 0$. In order that any vectors b_k^t may exist, the vector set a_i must evidently have a rank less than n, and conversely the rank of the vector set b_k^t must be less than n in order that any nonzero a_i vectors may fulfill the condition $a_i \cdot b_k^t = 0$.

If the null condition

$$a_i \cdot b_k^t = 0 \tag{126}$$

be compared with the system of homogeneous equations

$$a_i \cdot x = 0 \tag{127}$$

the discussion of the previous article yields the conclusion that if the rank of the matrix $\mathcal{A}$ is $r_a \leqq n$, then $n - r_a$ independent vectors b_k^t exist which fulfill the null condition, and hence the rank r_b of the matrix $\mathcal{B}$ may be as large as but no larger than $n - r_a$; that is, $r_b \leqq n - r_a$. Thus if

$$\mathcal{A} \times \mathcal{B} = 0 \tag{128}$$

then

$$r_a + r_b \leqq n \qquad [129]$$

in which $\mathcal{A}$ and $\mathcal{B}$ are square matrices of order n with the ranks r_a and r_b respectively.

It should be observed that the converse of this statement is not necessarily true, nor does it follow that the reversed product $\mathcal{B}\mathcal{A}$ is zero when $\mathcal{A}\mathcal{B}$ is zero.

An interesting application of this result is found in considering the product of a matrix $\mathcal{A}$ and its adjoint $\mathcal{A}^a$. From the definition of the adjoint matrix (see Art. 6, Ch. II) and determinant theory (see specifically Eqs. 45 and 46 of Ch. I), it follows that

$$\mathcal{A} \times \mathcal{A}^a = \mathcal{A}^a \times \mathcal{A} = \begin{bmatrix} A & 0 & 0 \cdots & 0 \\ 0 & A & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A \end{bmatrix} = A\mathcal{U} \qquad [130]$$

in which A is the determinant of $\mathcal{A}$. Hence if the rank of $\mathcal{A}$ is $(n - 1)$, $A = 0$, and according to the relations 128 and 129 the rank of $\mathcal{A}^a$ can be no greater than 1. Since the elements of $\mathcal{A}^a$ are the cofactors of $\mathcal{A}$, and at least one of these is not zero because the rank of $\mathcal{A}$ is $(n - 1)$, the rank of $\mathcal{A}^a$ cannot be less than 1. Hence the summary:

If $\mathcal{A}$ has the rank n, then $\mathcal{A}^a$ has the rank n.
If $\mathcal{A}$ has the rank $n - 1$, then $\mathcal{A}^a$ has the rank 1.
If $\mathcal{A}$ has any rank less than $n - 1$, then the rank of $\mathcal{A}^a$ is zero.

In connection with considerations of this sort it is useful to observe that if a given square matrix $\mathcal{A}$ has the rank 1, it possesses no more than one independent row and one independent column of elements, and hence it may be represented as the product of a column and a row matrix, thus:

$$\mathcal{A} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n \end{bmatrix} \times [\beta_1 \quad \beta_2 \cdots \beta_n] \qquad [131]$$

Similarly if a given square matrix has the rank 2, it possesses no more than two independent rows and two independent columns, and may

evidently be represented as

$$\mathcal{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} \end{bmatrix} \times \begin{bmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \beta_{21} & \cdots & \beta_{2n} \end{bmatrix} \qquad [132]$$

The generalization of these statements is obvious.

9. Sylvester's law of nullity

From the preceding considerations it follows that if a square matrix $\mathcal{A}$ of order n is given, and if this matrix has the rank r_a and the nullity p_a, a nonsingular matrix $\mathcal{B}$ can be found such that the product

$$\mathcal{C} = \mathcal{A} \times \mathcal{B} \qquad [133]$$

has p_a null columns. For example, if p_a of the columns of $\mathcal{B}$ are the components $b_{1k}, b_{2k}, \cdots b_{nk}$ of the vectors $b_k{}^t$ $(k = 1, 2, \cdots p_a)$, which are independent solutions to the homogeneous equations

$$\begin{aligned} a_1 \cdot b_k{}^t &= 0 \\ a_2 \cdot b_k{}^t &= 0 \\ \cdots&\cdots \\ a_n \cdot b_k{}^t &= 0 \end{aligned} \qquad [134]$$

p_a of the columns in $\mathcal{C}$ are composed of zeros. The remaining r_a columns in $\mathcal{B}$ may be chosen arbitrarily as long as they are independent, that is, as long as $\mathcal{B}$ turns out to be nonsingular.

There is no nonsingular matrix $\mathcal{B}$ for which $\mathcal{C}$ has more than p_a null columns, because (since $\mathcal{B}$ is nonsingular) $\mathcal{C}$ has the same rank as $\mathcal{A}$ (see Art. 9, Ch. II) and hence must have r_a independent columns. If, however, p_b of the remaining columns in $\mathcal{B}$ are then chosen to be linear combinations of the above p_a columns, $\mathcal{C}$ has $p_a + p_b$ null columns, and $\mathcal{B}$ is singular with nullity p_b. Moreover any matrix $\mathcal{B}$ with nullity p_b which produces $p_a + p_b$ null columns in $\mathcal{C}$ must have the form of the particular one chosen here, for $p_a + p_b$ of its columns must satisfy Eq. 134, and $n - p_b$ of its columns must be independent. In addition, it is not possible for any matrix $\mathcal{B}$ of rank r_b to produce more than $p_a + p_b$ null columns in $\mathcal{C}$, because any such new column requires that another column in $\mathcal{B}$ become a solution to Eq. 134, and thereby be a linear combination of p_a of the independent columns already in $\mathcal{B}$. The rank of $\mathcal{B}$ would then have to be less than r_b. The nullity of $\mathcal{C}$ is therefore at least $p_a + p_b$. But it cannot be greater than this because, if it were, post-multiplication of Eq. 133 by a nonsingular matrix $\mathcal{T}$ could produce

another null column in $\mathcal{C}$. This would demand that $\mathcal{B}\mathcal{T}$ be a matrix of rank r_b producing more than $p_a + p_b$ null columns in $\mathcal{C}$, which has been shown to be impossible. Hence no matrix $\mathcal{B}$ of nullity p_b can give $\mathcal{C}$ a nullity greater than $p_a + p_b$, that is,

$$p_c \leqq p_a + p_b \tag{135}$$

Of course $p_a + p_b$ may exceed n, in which case evidently p_c cannot be greater than n.

Now let it be supposed that two nonsingular transformation matrices $\mathcal{T}_A$ and $\mathcal{T}_B$ are determined such that

$$\mathcal{A}' = \mathcal{T}_A\mathcal{A} \tag{136}$$

is a matrix whose first r_a rows are independent and the remaining ones null, and

$$\mathcal{B}' = \mathcal{B}\mathcal{T}_B \tag{137}$$

is a matrix whose first r_b columns are independent and the remaining ones null. Then

$$\mathcal{T}_A\mathcal{A}\mathcal{B}\mathcal{T}_B = \mathcal{A}'\mathcal{B}' = \mathcal{T}_A\mathcal{C}\mathcal{T}_B = \mathcal{C}' \tag{138}$$

has the same rank as $\mathcal{C}$, but is evidently in the form

$$\mathcal{C}' = \begin{bmatrix} c_{11} & \cdots & c_{1r_b} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{r_a1} & \cdots & c_{r_ar_b} & 0 & \cdots & 0 \\ 0 & & \cdots & & & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & & \cdots & & & 0 \end{bmatrix} \tag{139}$$

It then becomes clear that the rank of $\mathcal{C}'$ and hence that of $\mathcal{C}$ cannot exceed r_a or r_b, whichever is the smaller. Or the nullity of $\mathcal{C}$ must be at least as great as p_a or p_b, whichever is the larger.

Together with Eq. 135 this conclusion yields the summary

$$p_r \leqq p_a + p_b \tag{140}$$
$$p_c \geqq \begin{Bmatrix} p_a > p_b \\ p_b > p_a \end{Bmatrix}$$

or stated in terms of rank,

$$r_c \geqq r_a + r_b - n \tag{141}$$
$$r_c \leqq \begin{Bmatrix} r_a < r_b \\ r_b < r_a \end{Bmatrix}$$

This useful statement regarding the rank or nullity of a product matrix in terms of the ranks or nullities of its components is known as *Sylvester's law of nullity* (or degeneracy).

10. Reduction of a square matrix to the diagonal form of its latent roots; normal co-ordinates

In view of the vector interpretation of the linear transformation

$$\begin{array}{c} a_{11}x_1 + \cdots + a_{1n}x_n = y_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n = y_n \end{array} \tag{142}$$

as discussed in Art. 3 of this chapter, it is interesting to inquire whether there exists a vector x whose transform y has the same direction in space, that is, whether a solution to Eqs. 142 exists which conforms to the stipulation

$$y = \lambda x \tag{143}$$

or more specifically

$$y_i = \lambda x_i \qquad (\text{for } i = 1, 2, \cdots n) \tag{144}$$

in which λ is a nonzero* factor.

Substituting the values given by Eq. 144 for the right-hand members of Eqs. 142, and transposing these to the left-hand sides, yields the conditions

$$\begin{array}{c} (a_{11} - \lambda)x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \cdots + a_{2n}x_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn} - \lambda)x_n = 0 \end{array} \tag{145}$$

This set of homogeneous equations has nontrivial solutions only when the determinant of its matrix (called the characteristic matrix of $\mathfrak{A}$) is zero, that is, when

$$\begin{vmatrix} (a_{11} - \lambda) & a_{12} & \cdots & a_{1n} \\ a_{21} & (a_{22} - \lambda) & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & (a_{nn} - \lambda) \end{vmatrix} = 0 \tag{146}$$

This determinant is called the *characteristic function* of the matrix $\mathfrak{A}$. It is evidently a polynomial in λ of the nth degree. The condition equation 146, which is called the *characteristic equation*, is an algebraic equation of the nth degree in λ, and hence there are n values of λ for which the matrix of Eqs. 145 has at most the rank $(n - 1)$. These so-called *latent roots* or *characteristic values* may be denoted by $\lambda_1, \lambda_2, \cdots \lambda_n$.

In the general case for which the matrix $\mathfrak{A}$ of the transformation 142 is unsymmetrical, the latent roots may be real or complex. Also they may

*The stipulation $\lambda = 0$ leads to the inquiry regarding the existence of solutions to the corresponding homogeneous equations, which has been discussed previously.

be distinct, or there may be coincident roots. If all the latent roots are distinct, the rank of the characteristic determinant for any particular latent root λ_s cannot be less than $(n-1)$; for if it were, all its $(n-1)$-rowed minors would be zero, and hence its derivative with respect to λ would also be zero when $\lambda = \lambda_s$. This fact would indicate that λ_s must be a multiple root of Eq. 146, which contradicts the assumption that the latent roots are distinct.

When the roots are distinct, one may conclude that there are n distinct directions for the vector x which remain unchanged by the transformation 142, that is, there are n directions for which the vector y coincides with the vector x. These are given by the n vectors $\overset{1}{x}, \overset{2}{x}, \cdots \overset{n}{x}$ which are found to be solutions to the n sets of homogeneous equations 145 resulting from substituting successively the characteristic values $\lambda_1, \lambda_2, \cdots \lambda_n$ for λ.

If the cofactors of the elements of the determinant in Eqs. 145 for the root $\lambda = \lambda_s$ are denoted by $\overset{s}{k}_{ik}$, the direction cosines of the vector $\overset{s}{x}$ are, according to Eq. 109, given by

$$\ell_{ks} = \frac{\overset{s}{k}_{ik}}{\sqrt{(\overset{s}{k}_{i1})^2 + (\overset{s}{k}_{i2})^2 + \cdots + (\overset{s}{k}_{in})^2}} \qquad (\text{for } k = 1, 2, \cdots n) \qquad [147]$$

in which the index i is arbitrary but must, of course, be the same for the determination of all the direction cosines corresponding to a given root λ_s.

The components $\overset{s}{x}_1, \overset{s}{x}_2, \cdots \overset{s}{x}_n$ of the vector $\overset{s}{x}$ may be taken equal to the cofactors $\overset{s}{k}_{i1}, \overset{s}{k}_{i2}, \cdots \overset{s}{k}_{in}$ respectively (i arbitrary) or to any multiple of these values since the length of the vector is not determined by Eqs. 145. Unit length for the vector $\overset{s}{x}$ results when its components $\overset{s}{x}_1, \overset{s}{x}_2, \cdots \overset{s}{x}_n$ are taken equal to the direction cosines $\ell_{1s}, \ell_{2s}, \cdots \ell_{ns}$ respectively.

With the aid of the column matrices

$$\overset{s}{x}] = \ell_s] = \begin{bmatrix} \overset{s}{x}_1 \\ \overset{s}{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \overset{s}{x}_n \end{bmatrix} = \begin{bmatrix} \ell_{1s} \\ \ell_{2s} \\ \cdot \\ \cdot \\ \cdot \\ \ell_{ns} \end{bmatrix} \qquad [148]$$

the transformation 142 for the condition 144, appropriate to the root $\lambda = \lambda_s$, may be expressed by the matrix equations

$$\mathcal{a}\overset{s}{x}] = \lambda_s \overset{s}{x}] \qquad [149]$$

or

$$\mathcal{a}\ell_s] = \lambda_s \ell_s] \qquad [150]$$

There are n matrix equations of this form for the n roots corresponding to the integer values 1, 2, $\cdots$ n for the indexes.

These n matrix equations may be combined into a single one by defining the matrix

$$\mathcal{L} = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1n} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix} \qquad [151]$$

in which the first column is the column matrix 148 for $s = 1$, the second column is the column matrix 148 for $s = 2$, and so forth. The set of n matrix equations 150, for $s = 1, 2, \cdots n$ is then given by

$$\mathcal{A}\mathcal{L} = \begin{bmatrix} \lambda_1\ell_{11} & \lambda_2\ell_{12} & \cdots & \lambda_n\ell_{1n} \\ \lambda_1\ell_{21} & \lambda_2\ell_{22} & \cdots & \lambda_n\ell_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_1\ell_{n1} & \lambda_2\ell_{n2} & \cdots & \lambda_n\ell_{nn} \end{bmatrix} \qquad [152]$$

or, by the introduction of the diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \cdots & 0 \\ 0 & \lambda_2 & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \qquad [153]$$

this is

$$\mathcal{A}\mathcal{L} = \mathcal{L}\Lambda \qquad [154]$$

It may now be shown that when the λ-roots are distinct the matrix $\mathcal{L}$ given by Eq. 151 is nonsingular or that the n vectors $\overset{1}{x}, \overset{2}{x}, \cdots \overset{n}{x}$ whose directions remain unchanged by the transformation 142 form a linearly independent set. The proof is readily given by assuming that this vector set is linearly dependent and showing that such an assumption leads to an absurdity.*

The assumption that the vector set $\overset{1}{x}, \overset{2}{x}, \cdots \overset{n}{x}$ is linearly dependent is equivalent to the statement that the vector relation

$$\gamma_1\overset{1}{x} + \gamma_2\overset{2}{x} + \cdots + \gamma_n\overset{n}{x} = 0 \qquad [155]$$

can be satisfied for γ-values other than all zeros. This vector relation is alternatively expressed by

$$\sum_{s=1}^{n} \gamma_s\overset{s}{x}] = 0 \qquad [156]$$

*The proof given here was suggested by Professor D. J. Struik of the Department of Mathematics at the Massachusetts Institute of Technology.

Multiplying Eq. 149 by γ_s and summing over s gives

$$\sum_{s=1}^{n} \mathcal{A}\gamma_s \overset{s}{x}] = \mathcal{A}\sum_{s=1}^{n} \gamma_s \overset{s}{x}] = \sum_{s=1}^{n} \gamma_s \lambda_s \overset{s}{x}] \qquad [157]$$

and in view of Eq. 156 this yields

$$\sum_{s=1}^{n} \gamma_s \lambda_s \overset{s}{x}] = 0 \qquad [158]$$

Repeating the same process but with Eq. 149 multiplied by λ_s in addition to γ_s, and this time using Eq. 158 instead of Eq. 156, gives

$$\sum_{s=1}^{n} \gamma_s \lambda_s^{2} \overset{s}{x}] = 0 \qquad [159]$$

In other words the assumption expressed by Eq. 155 or Eq. 156 together with the matrix Eq. 149 leads successively to relations of the form

$$\sum_{s=1}^{n} \gamma_s \lambda_s^{k} \overset{s}{x}] = 0 \qquad [160]$$

or in vector form

$$\gamma_1\lambda_1^{k}\overset{1}{x} + \gamma_2\lambda_2^{k}\overset{2}{x} + \cdots + \gamma_n\lambda_n^{k}\overset{n}{x} = 0 \qquad [161]$$

in which the exponent k can be 0, 1, 2, $\cdots$. For $k = 0$, Eq. 161 is the same as Eq. 155.

Since the λ-roots are assumed to be distinct, this result means that any number of independent relations of linear dependence among the vectors $\overset{1}{x}, \overset{2}{x}, \cdots \overset{n}{x}$ exist — clearly an absurdity. For example, in three-dimensional space, the existence of one relation of linear dependence means that the vectors lie in a plane; the existence of two independent relations of linear dependence means that the vectors lie in a straight line; the existence of three independent relations of linear dependence requires that the vectors be zero.

If two of the λ-roots are coincident, it might appear that two of the vectors $\overset{s}{x}$ become equal (or at most proportional), since they then represent two solutions to the same set of homogeneous equations 146. Under such conditions the modal matrix $\mathfrak{L}$, given by Eq. 151, would become singular by reason of the two proportional columns within it. It must be recognized, however, that the solution 147 for the various vectors $\overset{s}{x} = \ell_s$ is based upon the fact that the rank of the characteristic determinant is exactly $n - 1$, and although it is surely true when the λ-roots are distinct, it may or may not be valid in the case of repeated roots. In fact it is shown in Art. 4, Ch. IV, that, if the matrix $\mathcal{A}$ is symmetric, the occurrence of a repeated λ-root of order $p \leq n$ requires the rank of the characteristic determinant for that particular value of λ to be $n - p$. It is then possible

by the methods discussed in Art. 7 of the present chapter to find p independent solutions to this single set of homogeneous equations, and to make these solutions orthogonal and of unit length if desired. Under such conditions, the present proof of the independence of the solutions for all the λ-roots may be carried through without significant alteration, bearing in mind that the p solutions corresponding to the p equal λ-roots are already chosen to be independent of one another.

Unfortunately in the general case, when the matrix $\mathfrak{A}$ is not symmetric, there is no unique correspondence between the order of a repeated λ-root and the rank of the characteristic determinant. The nullity of the characteristic determinant may be less (but not greater) than the order of the repeated root, in which case solutions for n independent vectors $\overset{s}{x}$ cannot be found. For example, the unsymmetrical matrix

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

has a repeated latent root $\lambda = 1$ of order 3, whereas the nullity of the corresponding characteristic determinant is only 1. Since the rank of the latter determinant is $3 - 1 = 2$, it is possible to find only one independent vector $\overset{s}{x}$, and therefore any attempt to form a modal matrix yields a singular result. Hence the existence of a nonsingular modal matrix $\mathfrak{L}$ is guaranteed when and only when the matrix $\mathfrak{A}$ is either symmetric, or has n distinct latent roots.

In this case Eq. 154 may be premultiplied by $\mathfrak{L}^{-1}$ to give

$$\mathfrak{L}^{-1}\mathfrak{A}\mathfrak{L} = \Lambda \tag{162}$$

The matrix $\mathfrak{A}$ is thus reduced to the diagonal form of its latent roots.

The significance of this reduction, as far as the transformation 142 is concerned is recognized when the latter is written in matrix form,

$$\mathfrak{A}x] = y] \tag{163}$$

and the changes of variable expressed by

$$x] = \mathfrak{L}x'] \qquad \text{and} \qquad y] = \mathfrak{L}y'] \tag{164}$$

are considered. Then, in view of Eq. 162, the transformation 163 becomes

$$\Lambda x'] = y'] \tag{165}$$

which represents the equations

$$\begin{aligned} \lambda_1 x'_1 &= y'_1 \\ \lambda_2 x'_2 &= y'_2 \\ \cdots&\cdots\cdot \\ \lambda_n x'_n &= y'_n \end{aligned} \tag{166}$$

It thus becomes clear that the change of variable expressed by Eq. 164, which amounts to transforming from the co-ordinates x and y to new co-ordinates x' and y', reduces the set of simultaneous equations 142 to a set of n *distinct* equations, the solutions of which are immediately set down. Another way of looking at this situation is to say that in the new co-ordinate system to which x' and y' refer, the variables $x'_1 \cdots x'_n$ are no longer dependent upon each other as are the variables $x_1 \cdots x_n$ in the original co-ordinate system to which x and y refer.

In connection with physical problems, the new co-ordinates in which the equilibrium of the system appears to be expressed in terms of separate equations are called the *normal co-ordinates* of the system. The transformation to normal co-ordinates, which is effected by the matrix $\mathcal{L}$, apparently isolates the various modes of behavior of the physical system, and for this reason the matrix $\mathcal{L}$ is spoken of as the *modal matrix*.

In view of the fact that the vector $\overset{s}{x}$, which constitutes a solution to Eqs. 146, is determined in direction only, it may be noted that the matrix $\mathcal{A}$ is still reduced to the diagonal form of its latent roots if in the relation 162 the modal matrix $\mathcal{L}$ has its columns multiplied by arbitrary nonzero factors. In other words, the matrix $\mathcal{L}$ in Eq. 162 may be replaced by the product $\mathcal{L}\mathcal{D}$ in which $\mathcal{D}$ is an arbitrary nonsingular diagonal matrix. Since the inverse of $\mathcal{L}\mathcal{D}$ is $\mathcal{D}^{-1}\mathcal{L}^{-1}$, and $\mathcal{D}^{-1}\Lambda\mathcal{D} = \Lambda$, the truth of this statement is evident.

The type of transformation 162 to which the matrix $\mathcal{A}$ is subjected evidently corresponds to identical transformations for the variables, as shown by Eqs. 164. Thus a transformation of the matrix $\mathcal{A}$ of the form

$$\mathcal{C}^{-1}\mathcal{A}\mathcal{C} = \mathcal{B} \qquad [167]$$

corresponds to a transformation of the variables as expressed by

$$x] = \mathcal{C}x'] \qquad \text{and} \qquad y] = \mathcal{C}y'] \qquad [168]$$

in which $\mathcal{C}$ is necessarily nonsingular.

The geometrical significance of these relations can be studied through, first of all, recognizing that the transformation 142 may be looked upon as determining a point $y_1 \cdots y_n$ lying on a plane through the origin of co-ordinates defined by the equation

$$\beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_n y_n = 0 \qquad [169]$$

corresponding to a given point $x_1 \cdots x_n$ on a similar plane defined by the equation

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0 \qquad [170]$$

It may be helpful in this connection to consider the coefficients $\alpha_1 \cdots \alpha_n$ to be the components of a vector α. Then Eq. 170 states that $\alpha \cdot x = 0$,

which requires that the vector x be orthogonal to the vector α, and hence that the point $x_1 \cdots x_n$ lie in a plane normal to the direction of α. Equation 169 may similarly be interpreted, and it follows from the solution of simultaneous equations that the components of the corresponding vector β are given by the relation

$$\beta_k = a^*_k \cdot \alpha \qquad (\text{for } k = 1, 2, \cdots n) \qquad [171]$$

in which $a^*_1 \cdots a^*_n$ is the vector set of the matrix $\mathfrak{A}^*$ reciprocal to $\mathfrak{A}$. According to Eq. 74, the coefficients β_k are thus seen to be the contravariant components of α with respect to the oblique co-ordinate system defined by the vector set of $\mathfrak{A}$.

If the matrix $\mathfrak{A}$ is nonsingular, the transformation 142 yields a one-to-one correspondence between points on the two planes defined by Eqs. 169 and 170. One may say that Eqs. 142 transform planes into planes and hence also straight lines into straight lines.* For this reason a linear transformation is also spoken of as a *collineation*.

If the co-ordinates $x_1 \cdots x_n$ and $y_1 \cdots y_n$ define the points P and Q respectively, the collineation 142 relates any chosen points $P_1, P_2, \cdots$ (not necessarily on the same straight line) to corresponding points $Q_1, Q_2, \cdots$. The foregoing discussion in this article shows that a nonsingular collineation possesses n distinct points $P_1 \cdots P_n$ (called the *fixed points*) to which the corresponding points $Q_1 \cdots Q_n$ are so related that the directions OP_k and OQ_k (O denotes the origin of the co-ordinates) are coincident, and $OQ_k/OP_k = \lambda_k$. This so-called *descriptive* or *projective* property is evidently placed in evidence by the collineation which has the diagonal matrix Λ. The collineations with the matrices $\mathfrak{A}$ and Λ have the same projective properties, and are therefore spoken of as being *equivalent*.

More generally, any two collineations with the matrices $\mathfrak{A}$ and $\mathfrak{B}$ related by Eq. 167 have the same projective properties. Equation 167, which expresses the equivalence of the two collineations, is called a *collineatory transformation* of the matrix $\mathfrak{A}$.

The characteristic matrices of $\mathfrak{A}$ and $\mathfrak{B}$ are related by the same transformation. With the aid of the unit matrix $\mathfrak{U}$ of like order, the characteristic matrices of $\mathfrak{A}$ and $\mathfrak{B}$ are expressed by

$$(\mathfrak{A} - \lambda\mathfrak{U}) \qquad \text{and} \qquad (\mathfrak{B} - \lambda\mathfrak{U}) \qquad [172]$$

*The truth of the latter part of this statement may be seen from the fact that the transformation 142 relates points on a given continuous curve to points on another *continuous* curve. If a plane p_1 is transformed into a plane p'_1, and a second plane p_2 (not parallel to p_1) into the plane p'_2, points on the linear intersection of p_1 and p_2 must correspond to points on the linear intersection of p'_1 and p'_2; for these points are common to the planes p_1 and p_2 and, because of the continuity of the transformation, they must also be common to the planes p'_1 and p'_2.

But by Eq. 167

$$\mathcal{C}^{-1}(\mathcal{A} - \lambda\mathcal{U})\mathcal{C} = \mathcal{C}^{-1}\mathcal{A}\mathcal{C} - \lambda\mathcal{C}^{-1}\mathcal{U}\mathcal{C} = \mathcal{B} - \lambda\mathcal{U} \qquad [173]$$

The characteristic function of $\mathcal{B}$ (determinant of the characteristic matrix) is therefore

$$\begin{aligned} |\mathcal{B} = \lambda\mathcal{U}| &= |\mathcal{C}^{-1}(\mathcal{A} - \lambda\mathcal{U})\mathcal{C}| \\ &= C^{-1}|\mathcal{A} - \lambda\mathcal{U}|C = |\mathcal{A} - \lambda\mathcal{U}| \end{aligned} \qquad [174]$$

in which the bars enclosing the matrix expressions indicate that the determinant of the enclosed matrix is meant, and C is the determinant of the matrix $\mathcal{C}$.

The important conclusion to be drawn from this last relation is that the two matrices $\mathcal{A}$ and $\mathcal{B}$, which are connected by the transformation 167, have the same characteristic equation and hence have the same latent roots. In other words, *the latent roots of a matrix are invariant to a collineatory transformation.*

In this connection it is useful to note that Eqs. 162 and 167 yield

$$\mathcal{L}^{-1}\mathcal{C}\mathcal{B}\mathcal{C}^{-1}\mathcal{L} = \Lambda \qquad [175]$$

from which it is clear that the modal matrix for $\mathcal{B}$ is given by the product $\mathcal{C}^{-1}\mathcal{L}$.

11. The Cayley-Hamilton theorem

An interesting relationship follows from Eq. 162, namely

$$\mathcal{A} \times \mathcal{A} = \mathcal{A}^2 = \mathcal{L}\Lambda\mathcal{L}^{-1}\mathcal{L}\Lambda\mathcal{L}^{-1} = \mathcal{L}\Lambda^2\mathcal{L}^{-1} \qquad [176]$$

in which

$$\Lambda^2 = \Lambda \times \Lambda = \begin{bmatrix} \lambda_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & \lambda_n^2 \end{bmatrix} \qquad [177]$$

This relationship may readily be generalized with the result that the mth power of a nonsingular square matrix $\mathcal{A}$ is given by

$$\mathcal{A}^m = \mathcal{L}\Lambda^m\mathcal{L}^{-1} \qquad [178]$$

with

$$\Lambda^m = \begin{bmatrix} \lambda_1^m & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^m & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & \lambda_n^m \end{bmatrix} \qquad [179]$$

If $P(\lambda)$ is a polynomial in λ, that is,

$$P(\lambda) = a_p\lambda^p + a_{p-1}\lambda^{p-1} + \cdots + a_1\lambda + a_0 \tag{180}$$

the result expressed by Eq. 178 shows that*

$$\begin{aligned} P(\mathcal{A}) &= a_p\mathcal{A}^p + a_{p-1}\mathcal{A}^{p-1} + \cdots + a_1\mathcal{A} + a_0\mathcal{U} \\ &= \mathfrak{L}P(\Lambda)\mathfrak{L}^{-1} \end{aligned} \tag{181}$$

According to Eq. 179, however,

$$P(\Lambda) = \begin{bmatrix} P(\lambda_1) & 0 & 0 & \cdots & 0 \\ 0 & P(\lambda_2) & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & P(\lambda_n) \end{bmatrix} \tag{182}$$

and hence a polynomial in terms of a nonsingular matrix $\mathcal{A}$ becomes

$$P(\mathcal{A}) = \mathfrak{L}^{-1} \times \begin{bmatrix} P(\lambda_1) & 0 & 0 & \cdots & 0 \\ 0 & P(\lambda_2) & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & P(\lambda_n) \end{bmatrix} \times \mathfrak{L} \tag{183}$$

This result becomes particularly interesting when the polynomial $P(\lambda)$ is chosen to be the characteristic function of the matrix $\mathcal{A}$, that is,

$$P(\lambda) = |\mathcal{A} - \lambda\mathcal{U}| \tag{184}$$

Then $P(\lambda) = 0$ is the characteristic equation and hence $P(\lambda_1) = P(\lambda_2) = \cdots = P(\lambda_n) = 0$. The right-hand side of Eq. 183 is then zero, and the left-hand side is the characteristic function of $\mathcal{A}$ with the matrix $\mathcal{A}$ itself substituted in place of the variable λ. The conclusion follows that *the matrix $\mathcal{A}$ satisfies its own characteristic equation.*

The above derivation requires that the modal matrix $\mathfrak{L}$ be nonsingular, which means that the matrix $\mathcal{A}$ is either symmetric or has n distinct latent roots. In spite of these restrictions on the method of derivation employed here, it can be shown that the last result, which is known as the *Cayley-Hamilton theorem*, is true in general for any square matrix.†

A numerical illustration is given by the following:

$$\mathcal{A} = \begin{bmatrix} 0 & -1 & -3 \\ -1 & 2 & -1 \\ 1 & 1 & 4 \end{bmatrix} \tag{185}$$

*It should be observed that the constant term in the polynomial $P(\mathcal{A})$ is $a_0\mathcal{A}^0$, and that the zero power of a matrix is equal to the unit matrix of like order.

†See, for example, L. E. Dickson, *Modern Algebraic Theories* (New York: Benj. H. Sanborn and Co., 1926), p. 48.

The characteristic equation is given by

$$\begin{vmatrix} -\lambda & -1 & -3 \\ -1 & (2-\lambda) & -1 \\ 1 & 1 & (4-\lambda) \end{vmatrix} = 0$$

which yields

$$\lambda^3 - 6\lambda^2 + 11\lambda - 6 = 0 \tag{186}$$

or in factored form

$$(\lambda - 1)(\lambda - 2)(\lambda - 3) = 0 \tag{187}$$

Hence the Cayley-Hamilton theorem in this case states that

$$\mathcal{A}^3 - 6\mathcal{A}^2 + 11\mathcal{A} - 6\mathcal{U} = 0 \tag{188}$$

or in factored form that

$$(\mathcal{A} - \mathcal{U})(\mathcal{A} - 2\mathcal{U})(\mathcal{A} - 3\mathcal{U}) = 0 \tag{189}$$

Substituting the matrix $\mathcal{A}$ from Eq. 185 into the last equation yields the identity

$$\begin{bmatrix} -1 & -1 & -3 \\ -1 & 1 & -1 \\ 1 & 1 & 3 \end{bmatrix} \times \begin{bmatrix} -2 & -1 & -3 \\ -1 & 0 & -1 \\ 1 & 1 & 2 \end{bmatrix} \times \begin{bmatrix} -3 & -1 & -3 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \equiv 0 \tag{190}$$

which the reader may readily verify for himself.

It is worth noting from Eq. 188 that higher powers of the matrix $\mathcal{A}$ are expressible in terms of powers up to and including the $(n - 1)$th. Thus, for example,

$$\begin{aligned} \mathcal{A}^3 &= 6\mathcal{A}^2 - 11\mathcal{A} + 6\mathcal{U} \\ \mathcal{A}^4 &= 6\mathcal{A}^3 - 11\mathcal{A}^2 + 6\mathcal{A} = 25\mathcal{A}^2 - 60\mathcal{A} + 36\mathcal{U} \\ \mathcal{A}^5 &= 25\mathcal{A}^3 - 60\mathcal{A}^2 + 36\mathcal{A} = 90\mathcal{A}^2 - 239\mathcal{A} + 150\mathcal{U} \end{aligned} \tag{191}$$

and so forth.

In a similar manner, negative powers of $\mathcal{A}$ may be calculated. Thus from Eq. 188 in the above example

$$\begin{aligned} 6\mathcal{A}^{-1} &= \mathcal{A}^2 - 6\mathcal{A} + 11\mathcal{U} \\ 6\mathcal{A}^{-2} &= \mathcal{A} - 6\mathcal{U} + 11\mathcal{A}^{-1} = \tfrac{11}{6}\mathcal{A}^2 - 10\mathcal{A} + \tfrac{85}{6}\mathcal{U} \end{aligned} \tag{192}$$

and so forth. In particular, the inverse of $\mathcal{A}$ by this method is found to be

$$\mathcal{A}^{-1} = \tfrac{1}{6}\mathcal{A}^2 - \mathcal{A} + \tfrac{11}{6}\mathcal{U} = \tfrac{1}{6}\begin{bmatrix} 9 & 1 & 7 \\ 3 & 3 & 3 \\ -3 & -1 & -1 \end{bmatrix} \tag{193}$$

12. Symmetrical transformation

Returning to the discussion of Art. 10, note that when the matrix of the transformation 142 is symmetrical, the vectors $\overset{s}{x}$ whose directions are left unchanged by the transformation form a mutually orthogonal set.

The truth of this statement may be seen by considering Eq. 149 for two particular latent roots λ_s and λ_k, thus

$$\mathcal{A}\overset{s}{x}] = \lambda_s\overset{s}{x}] \tag{194}$$

$$\mathcal{A}\overset{k}{x}] = \lambda_k\overset{k}{x}] \tag{195}$$

The transpose of the matrices in Eq. 194, postmultiplied by $\overset{k}{x}]$ reads

$$\overset{s}{x}]_t\mathcal{A}_t\overset{k}{x}] = \lambda_s\overset{s}{x}]_t\overset{k}{x}] \tag{196}$$

and Eq. 195 premultiplied by the transpose of $\overset{s}{x}]$ is

$$\overset{s}{x}]_t\mathcal{A}\overset{k}{x}] = \lambda_k\overset{s}{x}]_t\overset{k}{x}] \tag{197}$$

Now if $\mathcal{A}$ is symmetrical, it is equal to its transpose $\mathcal{A}_t$, so that Eqs. 196 and 197 then yield

$$\lambda_s\overset{s}{x}]_t\overset{k}{x}] = \lambda_k\overset{s}{x}]_t\overset{k}{x}] \tag{198}$$

or

$$(\lambda_s - \lambda_k)\overset{s}{x}]_t\overset{k}{x}] = 0 \tag{199}$$

Written out, this reads

$$(\lambda_s - \lambda_k)[\overset{s}{x}_1\overset{s}{x}_2 \cdots \overset{s}{x}_n]\begin{bmatrix} \overset{k}{x}_1 \\ \overset{k}{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \overset{k}{x}_n \end{bmatrix} = (\lambda_s - \lambda_k)(\overset{s}{x}_1\overset{k}{x}_1 + \overset{s}{x}_2\overset{k}{x}_2 + \cdots + \overset{s}{x}_n\overset{k}{x}_n) = 0 \tag{200}$$

which is alternatively expressed by the scalar product

$$(\lambda_s - \lambda_k)\overset{s}{x} \cdot \overset{k}{x} = 0 \tag{201}$$

Since this result states that the scalar product of two vectors corresponding to two different latent roots is zero, the vectors must be normal to each other. In the case of repeated λ-roots, it is pointed out in Art. 10 of this chapter that the corresponding $\overset{s}{x}$ and $\overset{k}{x}$ can always be chosen to be orthogonal. Thus the apparent failure of the conclusion drawn from Eq. 201 in this instance is not significant.

In view of Eq. 201 it is also easy to show that the latent roots of a symmetrical matrix must be real. For if they were complex they would have to appear in pairs of conjugates,* for example,

$$\lambda_s = \alpha + j\beta \qquad \text{and} \qquad \lambda_k = \alpha - j\beta \tag{202}$$

*Since the matrix is assumed to have real coefficients.

and the vectors $\overset{s}{x}$ and $\overset{k}{x}$ satisfying Eqs. 194 and 195 would likewise have to be conjugate complex, thus

$$\overset{s}{x} = u + jv \qquad \text{and} \qquad \overset{k}{x} = u - jv \tag{203}$$

Their scalar product is given by

$$\overset{s}{x} \cdot \overset{k}{x} = u \cdot u + v \cdot v + j(u \cdot v - v \cdot u) \tag{204}$$

which is a real nonzero value. Hence Eq. 201 would demand that $\lambda_s = \lambda_k$ or

$$\alpha + j\beta = \alpha - j\beta \tag{205}$$

whence

$$\beta = 0 \tag{206}$$

In other words, the latent roots must be real.

When the vectors $\overset{s}{x}$ form a mutually orthogonal set, the modal matrix given by Eq. 151 becomes orthogonal, because its elements are then the direction cosines of a set of mutually orthogonal axes. Hence when the matrix $\mathcal{A}$ is symmetrical, its modal matrix has the property expressed by

$$\mathfrak{L}_t = \mathfrak{L}^{-1} \tag{207}$$

and the transformation of $\mathcal{A}$ to the diagonal form of its latent roots reads

$$\mathfrak{L}_t \mathcal{A} \mathfrak{L} = \Lambda \tag{208}$$

In this case, the normal co-ordinates of the physical system whose equilibrium is expressed by Eqs. 142 are orthogonal to each other.

Since the matrices encountered in practical problems are predominantly symmetrical, the results of the present article are quite significant. These matters are elaborated upon in Ch. IV.

PROBLEMS

1. Determine the rank of each of the following vector sets:

(a) $\begin{cases} (\ 5, 3, -1) \\ (10, 0, \ 2) \\ (\ 0, 1, \ 0) \end{cases}$ (b) $\begin{cases} (\ 3, 4, \ -1) \\ (-1.5, 2, -0.5) \\ (\ 6, 8, \ -2) \end{cases}$ (c) $\begin{cases} (1, 2, \ 4) \\ (1, 3, \ 9) \\ (1, 4, 16) \end{cases}$

(d) $\begin{cases} (\ 3a, \ 6b) \\ (-12a, 18b) \end{cases}$ (e) $\begin{cases} (1, 1, 1, 1) \\ (2, 2, 2, 2) \\ (3, 3, 3, 3) \\ (4, 4, 4, 4) \end{cases}$ (f) $\begin{cases} (1, 1, 1) \\ (1, e^{j(2\pi/3)}, e^{j(4\pi/3)}) \\ (1, e^{-j(2\pi/3)}, e^{-j(4\pi/3)}) \end{cases}$

(g) $\begin{cases} (1, 2^2, 2^3, \cdots 2^{n-1}) \\ (1, 3^2, 3^3, \cdots 3^{n-1}) \\ (1, 4^2, 4^3, \cdots 4^{n-1}) \\ \cdots\cdots\cdots\cdots\cdots \\ (1, n^2, n^3, \cdots n^{n-1}) \end{cases}$

2. A given vector set $a_1, a_2, \cdots a_m$ in an n-dimensional space (with $m < n$) is linearly dependent as expressed by

$$\alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_m a_m = 0$$

Show that all determinants of order m, formed from this vector set, vanish identically.

What can you say about the rank of the vector set if *all* the α's have finite nonzero values? If all but one of the α's are zero?

Which of the following vector sets is linearly dependent? What is the rank of each?

$$\begin{cases} (\ \ 6a, & -2b, & -4c, & 2) \\ (\ \ \ a, & b, & c, & d) \\ (-3a, & b, & 2c, & -1) \end{cases} \qquad \begin{cases} (6, 7, 9, 11, 13) \\ (2, 4, 6, \ 8, 10) \end{cases}$$

3. Consider the vector set $a_1, a_2, \cdots a_m$ in an n-dimensional space with $m = n + 1$, and suppose that the first n vectors $a_1, \cdots a_n$ are linearly independent. Show that one may write

$$a_m = \alpha_{11} a_1 + \alpha_{12} a_2 + \cdots + \alpha_{1n} a_n$$

and express the coefficients α_{1k} in terms of the determinant of the vector set $a_1, \cdots a_n$, its cofactors and the components of the mth vector.

Generalize this result for the case $m = n + p$ with $p > 1$, and illustrate with the following numerical example:

$$\begin{cases} (-1, & -2, & -3) \\ (\ \ 2, & 4, & 5) \\ (\ \ 1, & 1, & 1) \\ (\ \ 1, & 3, & 5) \\ (-1, & 0, & 1) \end{cases}$$

4. Consider three vectors in three-dimensional space. The vector a_1 has a length of eight units and is directed at an angle of 45° relative to the co-ordinate axis 3, while its projection upon the 1,2-plane makes angles of 45° with both axes 1 and 2. The vector a_2 lies in the 1,2-plane, having a length of six units, and oriented at angles of 30° and 60° respectively relative to the axes 1 and 2. The vector a_3 has a length of four units and coincides with axis 3. Write down the matrix corresponding to this vector set. Express each of these vectors linearly in terms of the three unit vectors coinciding with the co-ordinate axes.

5. The vector set $a_1, a_2, \cdots a_n$ in an n-dimensional space has the rank $m = n - 1$. Show that

$$\sum_{k=1}^{n} a_{ik} A_{nk} = 0 \quad \text{for } i = 1, 2, \cdots n$$

and interpret the significance of these equations geometrically. Suppose that this vector set is so chosen that the components a_{in} have finite nonzero values. A set of n vectors $b_1, b_2, \cdots b_n$ in an m-dimensional space is now defined with the components $b_{ik} = a_{ik}/a_{in}$. Obtain for this vector set the relations

$$\alpha_1 b_{i1} + \alpha_2 b_{i2} + \cdots + \alpha_m b_{im} = 1 \quad \text{for } i = 1, 2, \cdots n$$

giving the appropriate expressions for the coefficients α_k, and show that all these vectors terminate in the same plane. What is the equation of this plane?

Illustrate this problem for $n = 4$ with the vector set

$$\begin{cases} (\;\;1, & 1, 0, & 1) \\ (\;\;2, & 3, 1, & 1) \\ (\;\;4, & 0, 2, & 1) \\ (-2, & -2, 0, & -2) \end{cases}$$

6. Consider again the vector set of the previous problem and show that

$$\sum_{i=1}^{n} a_{ik}A_{in} = 0 \quad \text{for } k = 1, 2, \cdots n$$

Assuming that the vectors are so numbered that $A_{nn} \neq 0$, deduce the result

$$-a_n = \alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_{n-1} a_{n-1}$$

giving the values of the coefficients α_k. Determine this relation of linear dependence for the numerical vector sets:

$$\begin{cases} (\;\;1, & 2, & 1, 0) \\ (\;\;3, & 0, & 4, 1) \\ (\;\;\frac{1}{5}, & 0, & 0, 1) \\ (-2, & -4, & -2, 0) \end{cases} \qquad \begin{cases} (4, 2, 5, & 1) \\ (3, 0, 4, & 1) \\ (\frac{1}{8}, 0, 0, & -\frac{1}{8}) \\ (0, 3, 1, & 1) \end{cases}$$

7. Suppose the vector set $a_1 \cdots a_n$ of a matrix $\mathfrak{A}$ in an n-dimensional space has the rank $r = n - p$ $(1 < p < n)$, and assume the numbering of the vectors and components to be so chosen that the rth order minor in the upper left-hand corner of the determinant A is nonzero. Obtain linear relations between each of the p dependent vectors and r independent ones by the following procedure.

From the original matrix, form submatrices by striking out all but the first r columns and all but one of the last p rows. Call this row the $(r + i)$th. There are p such submatrices as i goes from $i = 1$ to $i = p$, and each submatrix then consists of r columns and $r + 1$ rows. A group of $r + 1$ minor determinants of order r is now obtained from each submatrix by deleting therefrom the kth row, with $k = 1, 2, \cdots r + 1$. Each of the minor determinants so generated is prefixed by a sign-controlling factor $(-1)^{(r+1+k)}$ to yield a "signed minor," denoted by α_{ik}. Note that $\alpha_{1,r+1} = \alpha_{2,r+1} = \cdots = \alpha_{p,r+1} \neq 0$.

Show next that the following systems of relations hold

$$\sum_{k=1}^{r+1} \alpha_{ik} a_{kq} = 0 \qquad \begin{matrix} (i = 1, 2, \cdots p) \\ (q = 1, 2, \cdots n) \end{matrix}$$

or that one may write the desired set of relations of linear dependence among the vectors:

$$\begin{aligned} \alpha_{11}a_1 + \alpha_{12}a_2 + \cdots + \alpha_{1,r+1}a_{r+1} &= 0 \\ \alpha_{21}a_1 + \alpha_{22}a_2 + \cdots + \alpha_{2,r+1}a_{r+2} &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ \alpha_{p1}a_1 + \alpha_{p2}a_2 + \cdots + \alpha_{p,r+1}a_{r+p} &= 0 \end{aligned}$$

Illustrate with the numerical vector set of rank 2:

$$\begin{cases} (2, 1, 6, 8) \\ (1, 1, 3, 5) \\ (0, 4, 0, 8) \\ (0, 1, 0, 2) \end{cases}$$

8. Consider two vector sets $a_1 \cdots a_n$ and $b_1 \cdots b_n$ with the square matrices $\mathfrak{A}$ and $\mathfrak{B}$.

(a) If the vectors in the two sets are connected by the relations

$$\sum_{k=1}^{n} c_{ik}a_k = b_i \quad \text{for } i = 1, 2, \cdots n$$

in which the matrix $\mathcal{C}$ is nonsingular, show that the components are connected by the matrix equation

$$\mathcal{C} \times \mathcal{A} = \mathcal{B}$$

and that the components of the corresponding transposed vector sets are related by

$$\mathcal{A}_t \times \mathcal{C}_t = \mathcal{B}_t$$

(b) Regard $a_1 \cdots a_n$ as a given vector set and $b_1 \cdots b_n$ as obtained from $a_1 \cdots a_n$ through linear transformation with the nonsingular matrix $\mathcal{C}$. Show that the rank of the vector set is invariant to this transformation but that such would not be the case if $\mathcal{C}$ were singular.

9. Two vector sets $a_1 \cdots a_n$ and $b_1 \cdots b_n$ with the square nonsingular matrices $\mathcal{A}$ and $\mathcal{B}$ have the matrix products

$$\mathcal{A}\mathcal{B} = \mathcal{P} \qquad \text{and} \qquad \mathcal{B}\mathcal{A} = \mathcal{Q}$$

(a) If the set $a_1 \cdots a_n$ is transformed by a nonsingular matrix $\mathcal{C}$ as indicated by

$$\mathcal{A}\mathcal{C} = \overline{\mathcal{A}}$$

how must the set $b_1 \cdots b_n$ be transformed in order that $\mathcal{P} = \overline{\mathcal{P}}$?

(b) Derive the corresponding relation between $\mathcal{Q}$ and $\overline{\mathcal{Q}}$, and show that the determinants of these matrices are equal.

10. Let P_1 and P_2 be two points in an n-dimensional orthogonal Cartesian co-ordinate system. The vectors x_1 and x_2 with components $x_{11} \cdots x_{1n}$ and $x_{21} \cdots x_{2n}$ respectively emanate from the origin of co-ordinates and terminate upon the points P_1 and P_2.

Express the direction cosines $\cos\alpha_1, \cos\alpha_2, \cdots \cos\alpha_n$ of a vector drawn from P_2 to P_1 in terms of the components of x_1 and x_2 and the distance l between P_1 and P_2, and prove the identity

$$\cos^2\alpha_1 + \cos^2\alpha_2 + \cdots + \cos^2\alpha_n \equiv 1$$

11. With respect to an oblique co-ordinate system, the contravariant components of a vector x are denoted by $\xi_1, \xi_2, \cdots \xi_n$ and its covariant components by $\xi^*_1, \xi^*_2, \cdots \xi^*_n$. The corresponding components of another vector y are $\eta_1, \eta_2, \cdots \eta_n$ and $\eta^*_1, \eta^*_2, \cdots \eta^*_n$. If the contravariant components of the two vectors are related by the matrix equation

$$[\xi_1 \cdots \xi_n] \times \mathcal{A} = [\eta_1 \cdots \eta_n]$$

and the covariant components by a similar equation with the matrix $\mathcal{B}$, how is $\mathcal{B}$ related to $\mathcal{A}$?

12. In an oblique co-ordinate system two vectors x_1 and x_2 have components which are conveniently expressed as the elements of the following row matrices:

$$[\xi_{11} \quad \xi_{12} \cdots \xi_{1n}] \qquad [\xi_{21} \quad \xi_{22} \cdots \xi_{2n}] \qquad \text{(contravariant)}$$

$$[\xi^*_{11} \quad \xi^*_{12} \cdots \xi^*_{1n}] \qquad [\xi^*_{21} \quad \xi^*_{22} \cdots \xi^*_{2n}] \qquad \text{(covariant)}$$

The lengths of these vectors are l_1 and l_2, θ is the angle between them, and d is the distance between their tips.

(a) Prove that:

$$d^2 = [\xi_{1k} - \xi_{2k}] \times [\xi^*_{1k} - \xi^*_{2k}]_t \qquad [1]$$

and

$$\cos\theta = [\lambda_{1k}] \times [\lambda^*_{2k}]_t = [\lambda_{2k}] \times [\lambda^*_{1k}]_t \qquad [2]$$

in which $\lambda_{sk} = \xi_{sk}/l_s$ and $\lambda^*_{sk} = \xi^*_{sk}/l_s$.

(b) Prove the invariance of these expressions under any linear transformation of co-ordinates.

(c) Show that Eq. 1 is equivalent to

$$d^2 = l_1^2 + l_2^2 - 2l_1l_2\cos\theta$$

(d) For any vector x with contravariant and covariant components ξ_k and ξ^*_k, show that

$$[\lambda_k] \times [\lambda^*_k]_t \equiv 1$$

Observe that $\lambda_k = \xi_k/l$ or $\lambda^*_k = \xi^*_k/l$ cannot be represented as the cosine of a real angle but that $-1 \leqq \lambda_k\lambda^*_k \leqq 1$. (The λ_k are called direction *parameters* and the λ^*_k the *moments* of the vector.)

(e) In a given n-dimensional co-ordinate system let λ_{sk} and λ^*_{sk} be the parameters and moments of a set of vectors x_s whose components are regarded as the elements of row matrices. If a new co-ordinate system is introduced through a transformation with the nonsingular matrix $[c_{ik}]$, the parameters and moments are carried over into the values $\bar{\lambda}_{sk}$ and $\bar{\lambda}^*_{sk}$. Show that the following expressions apply:

$$[\bar{\lambda}_{sk}] = [\lambda_{si}] \times [c_{ik}] \qquad \text{and} \qquad [\bar{\lambda}^*_{sk}] = [\lambda^*_{si}] \times [c_{ik}]_t^{-1}$$

13. The rows in the following matrix represent components of a given vector set in an orthogonal Cartesian co-ordinate system:

$$\begin{bmatrix} 4 & -5 & 1 \\ -2 & 5 & -1 \\ 3 & 0 & 1 \end{bmatrix}$$

(a) Compute the magnitudes of these vectors, their direction cosines, and the cosine of the angle between each pair.

(b) Find the inverse matrix and, for the inverse vector set which it represents, compute the corresponding quantities to those in part (a).

(c) Make a 120° isometric plot showing all vectors.

14. Make computations as in parts (a) and (b) of Prob. 13, considering the matrix

$$\begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{bmatrix}$$

Repeat the computations, using the transposed matrix.

15. (a) Consider the vector set of Prob. 13 to define the directions of an oblique co-ordinate system. With reference to the original orthogonal Cartesian system, the new vector [1, 3, −1] is given. For this vector compute the contravariant and covariant components, as well as its direction parameters and moments (see Prob. 12).

(b) Make an isometric plot showing the position of the vector and its abovementioned components.

(c) Repeat part (a) with reference to the vector set of Prob. 14.

16. Let the direction parameters λ_{1k} and λ_{2k} for $k = 1, 2, \cdots n$ be regarded as the contravariant components of two vectors v_1 and v_2 (sometimes called "versors"). If θ is the angle between these versors, show that

$$\sin^2\theta = \sum_{i,j,k,t=1}^{n} (g_{ij}g_{kt} - g_{ik}g_{jt})\lambda_{1i}\lambda_{1j}\lambda_{2k}\lambda_{2t}$$

$$= \left(\sum_{i,j=1}^{n} g_{ij}\lambda_{1i}\lambda_{1j}\right)\left(\sum_{k,t=1}^{n} g_{kt}\lambda_{2k}\lambda_{2t}\right) - \left(\sum_{i,k=1}^{n} g_{ik}\lambda_{1i}\lambda_{2k}\right)\left(\sum_{j,t=1}^{n} g_{jt}\lambda_{1j}\lambda_{2t}\right)$$

in which g_{ik} are the components of the fundamental metric tensor.

17. The matrix of the fundamental metric tensor in a given space of four dimensions is

$$\mathcal{G} = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 4 & 2 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Compute:

(a) The angles between the oblique axes.

(b) The corresponding metric coefficients (reciprocals of the lengths of the unit vectors $a_1 \cdots a_4$ of the oblique system).

(c) The length of a vector whose contravariant components are (5, 3, 2, 1), and its direction parameters.

(d) Compute the matrix $\mathcal{G}^*$.

(e) The angles between the oblique axes defined by $\mathcal{G}^*$.

(f) The corresponding metric coefficients.

(g) The covariant components and moments of the vector given in part (c).

18. Let $\mathcal{G}$ be the fundamental tensor corresponding to a system of oblique axes and let $\xi]$ be a column matrix whose elements are the contravariant components of a vector. Because of a co-ordinate transformation with the nonsingular matrix $\mathcal{C}$, the contravariant components of the given vector become the elements in the column matrix $\bar{\xi}] = \mathcal{C} \times \xi]$. Denoting by $\overline{\mathcal{G}}$ the matrix of the fundamental tensor with respect to the new co-ordinate axes, what are the expressions for $\overline{\mathcal{G}}$ and $\overline{\mathcal{G}}^*$ in terms of $\mathcal{G}$ and $\mathcal{G}^*$?

Taking for $\mathcal{G}$ the matrix in Prob. 17, and (5, 3, 2, 1) for the contravariant components of the given vector, compute the correspond:.ig components of the vector with respect to the new coordinates if the transformation matrix is

$$\mathcal{C} = \begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{bmatrix}$$

Compute the angles between the new co-ordinate axes.

19. Consider an oblique co-ordinate system and a point P such that all the contravariant components of a vector drawn from the origin to P are positive. The portions of the oblique axes coinciding with these vector components are regarded as the coterminous edges of a parallelepiped.

(a) Show that the volume of this parallelepiped is given by the formula

$$V = \sqrt{G}\,\xi_1\xi_2\xi_3$$

in which G is the determinant of the matrix $\mathfrak{G}$ and ξ_1, ξ_2, ξ_3 are the contravariant components of the vector.

(b) Write down the expressions for the areas of the faces of this parallelepiped which lie in the 1,2-plane, the 2,3-plane, and in the 3,1-plane respectively.

(c) Recognizing that the extension of the above formula to more dimensions reads

$$V = \sqrt{G}\,\xi_1\xi_2 \cdots \xi_n$$

compute the volume of the hyper-parallelepiped in four-dimensional space defined by the matrix $\mathfrak{G}$ of Prob. 17 and a vector with the contravariant components [1, 2, 3, 1].

Compute the area of the parallelogram for each pair of co-ordinates.

20. Let $x_1, x_2, \cdots x_n$ be the contravariant co-ordinates of a point P in n-dimensional space with respect to a set of oblique axes. This system is regarded as immersed in an m-dimensional space with $m > n$. In the m-dimensional space the co-ordinates of the same point P with respect to an orthogonal set of axes are $y_1, y_2, \cdots y_m$. The orthogonal and the oblique axes have coincident origins.

Show that the y_k's are expressible in terms of the x_k's in the form

$$\sum_{k=1}^{n} a_{ik}x_k = y_i \qquad (i = 1, 2, \cdots m)$$

but that not more than n of the y_k's are independent. More specifically if the first n of the above equations are independent show that $y_{n+1}, y_{n+2}, \cdots y_m$ may be expressed in terms of $y_1 \cdots y_n$ by means of the relations

$$\frac{1}{A}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ik}A_{jk}y_j = y_i \qquad (i = n + 1, n + 2, \cdots m)$$

in which A is the determinant corresponding to the first n rows of the matrix $[a_{mn}]$ and A_{jk} are the appropriate cofactors.

Show further that the squared line element $\overline{ds}^2$ is expressed by

$$\overline{ds}^2 = \sum_{k=1}^{m} (dy_k)^2 = \sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial y_i}{\partial x_k}\frac{\partial y_i}{\partial x_j}\,dx_k\,dx_j$$
$$= \sum_{k,j=1}^{n} g_{kj}dx_k\,dx_j$$

in which

$$g_{kj} = \sum_{i=1}^{m} \frac{\partial y_i}{\partial x_k}\frac{\partial y_i}{\partial x_j}$$

are the components of the fundamental metric tensor in the n-dimensional space.

Compute the values of g_{kj} in terms of the a_{rs} of the transformation.

21. Let $x]$ and $\alpha]$ be column matrices and regard their elements $x_1 \cdots x_n$ and $\alpha_1 \cdots \alpha_n$ as components of a variable vector x and a fixed vector α with respect to an orthogonal co-ordinate system.

(a) Show that the equation

$$\alpha]_t x] = d \qquad \text{(a scalar)}$$

represents the equation of a plane, that the vector α is normal to this plane, and that its distance from the origin is $l = d/(\text{magnitude of } \alpha)$.

(b) In transforming to a new system of oblique co-ordinate axes, show that the distance of l and the point defined by α will not change if the vector x is subjected to

the transformation

$$\mathcal{A}^* x] = \xi]$$

whereas the vector α transforms as

$$\mathcal{A}\alpha] = \alpha^*]$$

22. Let $\xi]$ and $\alpha]$ be column matrices whose elements are the contravariant components of a variable vector and a fixed vector respectively with reference to an oblique set of axes in three-dimensional space. Show that the plane defined by

$$\alpha]_t \times \xi] = d \qquad \text{(a scalar)}$$

cuts the co-ordinate axes at points whose distances to the origin are given by

$$s_j = \sqrt{g_{jj}}\frac{d}{\alpha_j}$$

and that the cosines of the angles that the plane makes with the co-ordinate axes are

$$\cos\theta_j = \frac{\alpha_j\sqrt{g_{jj}}}{|\alpha|}$$

in which $|\alpha|$ denotes the length of the vector α.

23. Let the equations of two planes be given by

$$\alpha]_t \times \xi] = d_1 \qquad \text{and} \qquad \beta]_t \times \xi] = d_2$$

Show that they intersect at an angle θ for which

$$\cos\theta = \frac{1}{|\alpha|\,|\beta|}\alpha]_t \times \mathcal{G} \times \beta]$$

and compute this angle for the planes corresponding to vectors α and β with components 3, -1, 4 and 2, 8, 6 with

$$\mathcal{G} = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

24. Suppose that a space of three dimensions is immersed in a space of four dimensions. Let y be a vector whose components when referred to a set of orthogonal axes in the four-dimensional space are y_1, y_2, y_3, y_4. The point determined by this vector lies also in the three-dimensional space and is there characterized by the co-ordinate x_1, x_2, x_3.

Denoting ree constant vectors in the four-dimensional space by α, β, γ, show that the equation

$$\begin{vmatrix} y_1 & y_2 & y_3 & y_4 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \end{vmatrix} = d \qquad \text{(a scalar)}$$

represents a plane in the three-dimensional space and find the equation of this plane in terms of the x-co-ordinates, recognizing that one may write

$$\sum_{k=1}^{3} a_{ik}x_k = y_i \quad \text{for} \quad i = 1, 2, 3, 4$$

25. With respect to an orthogonal set of axes in three-dimensional space, three points P_1, P_2, P_3 determined by the linearly independent vectors α, β, λ lie in a plane.

(a) Write the expression determining a variable vector x which terminates in this plane.

(b) Determine a unit vector perpendicular to the plane.

(c) If the points P_1, P_2, P_3 are transformed into $\bar{P}_1$, $\bar{P}_2$, $\bar{P}_3$ through a nonsingular transformation with the matrix $\mathfrak{A}$ such that the vectors become

$$\bar{\alpha}] = \mathfrak{A}_t\alpha]$$

$$\bar{\beta}] = \mathfrak{A}_t\beta]$$

$$\bar{\gamma}] = \mathfrak{A}_t\gamma]$$

show that the transformed vector $\bar{x}] = \mathfrak{A}_t x]$ terminates in the new plane.

(d) Let the components of α, β, γ be given by $(3, 2, 1)$, $(4, 1, 5)$, $(1, 1, 1)$, and find the transformed plane corresponding to

$$\mathfrak{A} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 1 \end{bmatrix}$$

26. Let P be a variable point on the surface of a hypersphere of radius ρ and with its center at P_0. P and P_0 are determined respectively by the variable vector x and fixed vector α. The co-ordinate system is orthogonal and the equation of the hypersphere is given by

$$[x_k - \alpha_k]_t \times [x_k - \alpha_k] = \rho^2$$

in which $x_1 \cdots x_n$ and $\alpha_1 \cdots \alpha_n$ are regarded as elements of column matrices.

If the points P and P_0 are transformed into $\bar{P}$ and $\bar{P}_0$ through a nonsingular transformation with the matrix $\mathfrak{A}$, compute the distance between the two new points $\bar{P}$ and $\bar{P}_0$ and show that the new point $\bar{P}$ does not in general lie upon a new hypersphere.

Illustrate for $n = 2$ with $\alpha_1 = \alpha_2 = 1$ and for the matrix

$$\mathfrak{A} = \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}$$

find the locus of the point $\bar{P}$ as the point P traverses its circle.

27. (a) If a matrix $\mathfrak{B}$ is obtained through a collineatory transformation of the matrix $\mathfrak{A}$, show that the matrices $\mathfrak{B}^n$ and $\mathfrak{B}^{-1}$ are obtained through the same collineatory transformation of the matrices $\mathfrak{A}^n$ and $\mathfrak{A}^{-1}$.

(b) If matrices $\mathfrak{B}$ and $\mathfrak{S}$ are obtained through the same collineatory transformation of matrices $\mathfrak{A}$ and $\mathfrak{D}$ respectively, show that the matrices $(\mathfrak{B} + \mathfrak{S})$, $\mathfrak{B}\mathfrak{S}$, $\mathfrak{S}\mathfrak{B}$, $(\mathfrak{B}\mathfrak{S})^{-1}$, and $(\mathfrak{S}\mathfrak{B})^{-1}$ are obtained through the same collineatory transformation of $(\mathfrak{A} + \mathfrak{D})$, $\mathfrak{A}\mathfrak{D}$, $\mathfrak{D}\mathfrak{A}$, $(\mathfrak{A}\mathfrak{D})^{-1}$, and $(\mathfrak{D}\mathfrak{A})^{-1}$ respectively.

(c) If $\mathfrak{B}$ is obtained through a collineatory transformation of $\mathfrak{A}$, how is $\mathfrak{B}_t$ related to $\mathfrak{A}_t$?

(d) Make a collineatory transformation of the matrix

$$\mathfrak{A} = \begin{bmatrix} 3 & 2 & 1 \\ -4 & 1 & 1 \\ -1 & -2 & 5 \end{bmatrix} \quad \text{with the matrix } \mathfrak{C} = \begin{bmatrix} -1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

(e) Let x and α respectively be a variable and a fixed vector, and assume that $\alpha]_t \times x] = 0$. If $y] = \mathfrak{C}^{-1}\mathfrak{A}\mathfrak{C}x]$, in which $\mathfrak{A}$ and $\mathfrak{C}$ are nonsingular, determine the

vector β satisfying the relation $\beta]_t y] = 0$ and illustrate for $n = 3$ with $\alpha]_t = [3, 2, 1]$ and matrices $\mathfrak{A}$ and $\mathfrak{C}$ as given in part (d).

28. Find the latent roots of the following matrices

$$\begin{bmatrix} 1 & -13 & 24 \\ 0 & -11 & 21 \\ 0 & -6 & 12 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \quad \begin{bmatrix} 191 & -257 & 16 \\ 114 & -152 & 8 \\ 85 & -115 & 8 \end{bmatrix} \quad \begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}$$

29. Find the latent roots of the matrix

$$\begin{bmatrix} 2 & 1 & 1 \\ -2 & 1 & 3 \\ 3 & 1 & -1 \end{bmatrix}$$

Form the modal matrix $\mathfrak{L}$ and check numerically the relation $\mathfrak{L}^{-1}\mathfrak{A}\mathfrak{L} = \Lambda$. Make an isometric plot showing the directions of the principal axes.

30. Prove that the determinant and the rank of a matrix remain invariant under a collineatory transformation.

31. Show that the roots of the characteristic function are invariant to elementary transformations of the characteristic matrix. Through an appropriate succession of elementary transformations of the characteristic matrix, show that the latent roots of the matrix

$$\begin{bmatrix} 5 & 0 & -1 & 0 \\ 0 & 5 & 0 & -1 \\ 9 & 1 & 5 & 0 \\ 0 & 9 & 0 & 5 \end{bmatrix}$$

are $\lambda_1 = \lambda_3 = 5 - j3$ and $\lambda_2 = \lambda_4 = 5 + j3$.

CHAPTER IV

Quadratic Forms

1. THE QUADRATIC FORM ASSOCIATED WITH A LINEAR TRANSFORMATION

If in the linear transformation

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned} \qquad [1]$$

the equations are multiplied respectively by $x_1, x_2, \cdots x_n$, the sum of the resulting left-hand members is a rational entire function which is homogeneous and quadratic in the variables $x_1 \cdots x_n$. This function is called a *quadratic form*. More specifically it is referred to as the quadratic form associated with the linear transformation 1.

Written out, this function has the appearance:

$$\begin{aligned} F = \quad & a_{11}x_1^2 + a_{12}x_2x_1 + \cdots + a_{1n}x_nx_1 \\ & + a_{21}x_1x_2 + a_{22}x_2^2 + \cdots + a_{2n}x_nx_2 \\ & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ & + a_{n1}x_1x_n + a_{n2}x_2x_n + \cdots + a_{nn}x_n^2 \end{aligned} \qquad [2]$$

The terms on the principal diagonal of this square array involve the squares of the variables $x_1 \cdots x_n$; the remaining terms involve all the possible cross-products.

The matrix $\mathfrak{A}$ of the transformation 1 is called the matrix of the quadratic form, and its determinant A is referred to as the *discriminant* of the quadratic form associated with this transformation. This chapter is concerned with *real* quadratic forms, that is, such forms which have matrices with real elements.

The writing of this function can be abbreviated through the use of a double summation, thus

$$F = \sum_{i,\,k=1}^{n} a_{ik}x_ix_k \qquad [3]$$

Here the terms for $i = 1$ and k from 1 to n are those in the first row of Eq. 2; the terms for $i = 2$ and k from 1 to n are those in the second row of Eq. 2, and so forth.

Since the terms with a_{ik} and a_{ki} for like values of i and k $(k \neq i)$ involve the product of the same pair of variables x_i and x_k (for example, the terms

$a_{12}x_1x_2$ and $a_{21}x_2x_1$), the quadratic form is evidently no less general if the matrix $\mathfrak{A}$ of the associated linear transformation is assumed to be symmetrical; that is, if

$$a_{ik} = a_{ki} \qquad \text{or} \qquad \mathfrak{A}_t = \mathfrak{A} \tag{4}$$

Hence in the discussion of quadratic forms it is possible to assume that the associated linear transformation is a symmetrical one, and thus to make available to the discussion the special properties of such transformations pointed out in the last article of the preceding chapter.

A quadratic form may also be thought of as obtained through the squaring of a linear form; thus

$$F = (\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n)^2 \tag{5}$$

although the result is less general because it contains only n arbitrary coefficients.* Its formal appearance is, however, the same, as may be seen if all the terms in the full square represented by Eq. 5 are systematically written down; thus

$$\begin{aligned} F = \quad & \alpha_1{}^2x_1{}^2 + \alpha_1\alpha_2x_1x_2 + \cdots + \alpha_1\alpha_nx_1x_n \\ & + \alpha_2\alpha_1x_2x_1 + \alpha_2{}^2x_2{}^2 + \cdots + \alpha_2\alpha_nx_2x_n \\ & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ & + \alpha_n\alpha_1x_nx_1 + \alpha_n\alpha_2x_nx_2 + \cdots + \alpha_n{}^2x_n{}^2 \end{aligned} \tag{6}$$

Identification of Eqs. 2 and 6 leads to the relation

$$a_{ik} = \alpha_i\alpha_k \tag{7}$$

The function F as given by Eq. 5 is more convenient for the derivation of the following useful relations. Thus it is readily seen that the partial derivative of F with respect to one of the variables x_i is given by

$$\frac{\partial F}{\partial x_i} = 2\alpha_i(\alpha_1x_1 + \alpha_2x_2 + \cdots + \alpha_nx_n) \tag{8}$$

By substitution of Eq. 7, the corresponding result for the function F of Eq. 2 reads†

$$\frac{1}{2}\frac{\partial F}{\partial x_i} = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \tag{9}$$

This result allows the quadratic form to be expressed in terms of its par-

*The rank of the quadratic form (which is the rank of its matrix) is in this case 1, as may be seen from the fact that the coefficients given by Eq. 7 are formed in the same fashion as are those of the matrix in Eq. 131, Ch. III. For the present discussion the rank of the quadratic form is immaterial.

†This result may, of course, also be obtained directly from Eq. 2, utilizing the symmetry condition 4.

tial derivatives thus,

$$F = \frac{1}{2}\left\{\frac{\partial F}{\partial x_1}\cdot x_1 + \frac{\partial F}{\partial x_2}\cdot x_2 + \cdots + \frac{\partial F}{\partial x_n}\cdot x_n\right\} \tag{10}$$

or in the more compact notation as

$$F = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial F}{\partial x_i}x_i \tag{11}$$

Correspondingly, the linear transformation 1 associated with a quadratic form F may, according to Eq. 9, be written in the form

$$\frac{1}{2}\frac{\partial F}{\partial x_i} = y_i \qquad (\text{for } i = 1, 2, \cdots n) \tag{12}$$

2. Geometrical interpretation of a quadratic form; the quadric surface associated with a linear transformation

According to the derivation of the quadratic form from its associated linear transformation 1, it is clear that the function F may be identified with the following *bilinear* form in the variables $x_1 \cdots x_n$ and $y_1 \cdots y_n$:

$$F = x_1y_1 + x_2y_2 + \cdots + x_ny_n \tag{13}$$

If $x_1 \cdots x_n$ and $y_1 \cdots y_n$ are regarded as the components of the vectors x and y respectively, the quadratic form is represented by the scalar product

$$F = x \cdot y \tag{14}$$

That is to say, the value of the quadratic form may be expressed as the scalar product of the vector x with its transform y. Of interest in connection with this interpretation is inquiry into the significance of the equation

$$F = 1 \tag{15}$$

or in other words into the question: For what values of the variables $x_1 \cdots x_n$ does the quadratic form maintain a constant value? In Eq. 15, the constant value is arbitrarily set equal to unity. The question may, in view of Eq. 14, be stated in another way: For what lengths of the vector x, which, it is assumed, may have all possible orientations, does the scalar product of the vector with its transform y have the fixed value unity?

As the vector x assumes all possible orientations, under the stipulation that its scalar product with the vector y maintains a unit value, its length is evidently forced to vary in a very definite manner. The tip of the vector x, therefore, describes a surface (in n-dimensional space this surface is $n - 1$ dimensional), and the solution to the inquiry stated above is seen

to resolve itself into the determination of the surface defined by Eq. 15, in which F is a function of the variables $x_1 \cdots x_n$.

Since the function F is quadratic in the variables $x_1 \cdots x_n$, the surface in question is evidently one of second order, which is also known as a *quadric surface*. In three dimensions the quadric surfaces are those of the familiar ellipsoid, the paraboloid, or the hyperboloid. The quadric surfaces in n-dimensional space may be visualized in a similar manner, although their reality is, of course, lost.

Since the function F contains no linear terms, its value is unchanged if the algebraic signs of all the variables are reversed. This fact means that the vector x satisfying Eq. 15 has the same length when its direction is reversed. Hence the quadric surface is symmetrical about the origin. A surface of this type is referred to as a *central* quadric surface (alternatively as a *central conic*).

The central quadric surface has hyperbolic or ellipsoidal characteristics, or both, depending upon the values and the algebraic signs of the coefficients a_{ik} in the form F. This question is discussed further in a subsequent article.

3. Transformation of Variables

When the variables $x_1 \cdots x_n$ in the quadratic form given by Eq. 3 are subjected to a linear transformation such as

$$x] = \mathcal{C}x'] \qquad [16]$$

in which $x]$ and $x']$ are the column matrices

$$x] = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad x'] = \begin{bmatrix} x'_1 \\ x'_2 \\ \cdot \\ \cdot \\ \cdot \\ x'_n \end{bmatrix} \qquad [17]$$

and $\mathcal{C}$ is an arbitrary square matrix of order n, then F becomes a quadratic form in the new variables $x'_1 \cdots x'_n$. When the matrix $\mathcal{C}$ is nonsingular, the new variables are uniquely related to the original variables $x_1 \cdots x_n$, and hence for given variables $x_1 \cdots x_n$ the quadratic form expressed in terms of the new variables has the same value. In other words, the quadratic form in the original variables is identical with that expressed in terms of the new variables, the change of variable being nothing more than a formal change in the notation.

At the same time, the transformation expressed by Eq. 16 may have the geometrical significance of a change of co-ordinate systems. In any

case, the immediate problem is to determine the matrix of the quadratic form for the new variables in terms of its matrix for the original variables and that of the transformation 16.

For this purpose, it is useful to recognize that Eq. 2 for the quadratic form may be written in matrix form. First the transformation 1 is written in the matrix form

$$\mathcal{A}x] = y] \qquad [18]$$

in which $\mathcal{A}$ is the matrix of the coefficients appearing in Eqs. 1, and $y]$ is the column matrix

$$y] = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad [19]$$

According to the form for F given by Eq. 13, it is seen that

$$F = \lfloor x \rfloor y] \qquad [20]$$

in which $\lfloor x \rfloor$ is the transpose of the column matrix $x]$. Substituting for $y]$ from Eq. 18 yields the desired matrix expression for F, namely,

$$F = \lfloor x \rfloor \mathcal{A}x] \qquad [21]$$

The expression for F in terms of the new variables $x'_1 \cdots x'_n$ is now readily obtained by substituting for $x]$ from Eq. 16, and recognizing that the transpose of this equation gives

$$\lfloor x \rfloor = \lfloor x' \rfloor \mathcal{C}_t \qquad [22]$$

Thus it is found that

$$F = \lfloor x' \rfloor \mathcal{C}_t \times \mathcal{A} \times \mathcal{C}x'] \qquad [23]$$

which may be written

$$F = \lfloor x' \rfloor \mathcal{B}x'] \qquad [24]$$

in which

$$\mathcal{B} = \mathcal{C}_t \times \mathcal{A} \times \mathcal{C} \qquad [25]$$

is the desired matrix of F for the new variables x'_s.

Equation 25 gives the transformation formula for the matrix of a quadratic form when its variables are subjected to a linear transformation with the matrix $\mathcal{C}$. This is called a *congruent* transformation of $\mathcal{A}$. It is recognized that since $\mathcal{A}$ is symmetrical, the matrix $\mathcal{B}$ is also symmetrical.

It may be noted that the transformation expressed by Eq. 25 is very similar to the collineatory transformation given by Eq. 167 of Ch. III. The two transformations become identical if the matrix $\mathcal{C}$ is an orthogonal one, that is, if the transformation 16 relates the co-ordinates $x_1 \cdots x_n$ in a given rectangular co-ordinate system to the co-ordinates $x'_1 \cdots x'_n$ in another rectangular co-ordinate system with the same origin. This conclusion is recognized from the fact that for such a transformation $\mathcal{C}_t = \mathcal{C}^{-1}$ (orthogonality condition).

Moreover, since the determinant of an orthogonal matrix has the value ± 1 (see Art. 6, Ch. II, and Art. 4, Ch. III) it is clear that for this kind of co-ordinate transformation the determinant of the matrix $\mathcal{B}$ is equal to that of $\mathcal{A}$. In other words, *the discriminant of a quadratic form is invariant to an orthogonal transformation of its variables.*

4. THE PRINCIPAL AXES OF A QUADRIC SURFACE; THE REDUCTION OF A QUADRATIC FORM TO A SUM OF SQUARES; DEGENERACY AND RANK

The quadric surface defined by Eq. 15, although central, does not have its principal axes (like the major and minor axes of an ellipse) coincident with the axes of the reference co-ordinate system (the set of rectangular axes on which the components $x_1 \cdots x_n$ of the vector x are projected). If they were coincident, only the square terms in Eq. 2 would be present. The presence of the cross-product terms means that the principal axes of the quadric surface have arbitrary orientations in space relative to the directions of the reference axes.

It is a common problem in analytic geometry to determine the orientations of the axes of the quadric surface defined by Eq. 15 when the function F is given in the form of Eq. 2. If these directions can be found and chosen as the axes of a new co-ordinate system, a change of variables, which amounts to a transformation to these new co-ordinates, should have the effect of eliminating the cross-product terms from the expression for F.

Thus the geometrical problem of finding the principal axes of a quadric surface is seen to be essentially the same as the algebraic problem of reducing an arbitrary quadratic form to a sum of squares.

According to the vector interpretation given in Art. 2 of this chapter, the quadric surface is generated by all possible x-vectors whose scalar products with their transforms y have the fixed value unity. For any given direction of the vector x, its length is the distance from the origin to the surface. If the direction is along a principal axis, this distance is evidently a maximum or a minimum as compared with the distances for

any neighboring directions.* This fact forms a convenient basis for finding the principal axes, since it reduces the task to a maximum-minimum problem.

The square of the length of the vector x is given by the expression

$$|x|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \tag{26}$$

Since the maxima and minima of the length of x are coincident with those of the square of this length, there is no need to take the square root of the expression 26. In order that this length may represent the distance from the origin to the quadric surface, however, the variables $x_1 \cdots x_n$ must satisfy the equation

$$F = 1 \tag{27}$$

in which F is given by Eq. 2.

According to the usual procedure for finding maxima, the expression for one of the variables obtained from Eq. 27 is to be substituted into Eq. 26, and the resultant function made a maximum or minimum by setting its partial derivatives with respect to the remaining variables equal to zero. Obviously, this undertaking is exceedingly awkward in view of the complicated form of the function F. Hence in problems of this sort a slightly altered procedure, known as the method of determining *conditioned* maxima (or minima), is adopted.

Thus it is said that the function 26 is to be made a maximum subject to the condition imposed by the auxiliary relation 27. The procedure is to form the function

$$f(x_1, \cdots x_n) = x_1^2 + x_2^2 + \cdots + x_n^2 - \mathrm{H}(F - 1) \tag{28}$$

in which H is an arbitrary constant (called a Lagrangian multiplier). In view of Eq. 27, this function is evidently the same as $|x|^2$, or rather it should be said that the function 28 becomes the same as $|x|^2$ when the condition 27 is fulfilled.

The partial derivatives of $f(x_1, \cdots x_n)$ equated to zero, that is

$$\frac{\partial f}{\partial x_k} = 0 \quad \text{for } k = 1, 2, \cdots n \tag{29}$$

yield n equations. Together with the condition 27, these may be solved for the multiplier H and the special values of $x_1 \cdots x_n$ which yield the desired conditioned maximum.

*This circumstance is readily visualized in the case of an ellipsoid, for which all these distances are real. For the moment, the geometrical interpretation may be confined to this type of surface, although the results of the present discussion, as appears later on, are not restricted to ellipsoidal quadric surfaces. The discussion in Art. 10 deals further with this question.

Forming the partial derivatives in Eq. 29 for the function expressed by Eq. 28 gives

$$
\begin{aligned}
2x_1 - \mathrm{H}\frac{\partial F}{\partial x_1} &= 0 \\
2x_2 - \mathrm{H}\frac{\partial F}{\partial x_2} &= 0 \\
\cdots\cdots\cdots&\cdots \\
2x_n - \mathrm{H}\frac{\partial F}{\partial x_n} &= 0
\end{aligned}
\qquad [30]
$$

If the values for the partial derivatives of F given by Eq. 9 are substituted, these equations become

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= \lambda x_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= \lambda x_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots&\cdots \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= \lambda x_n
\end{aligned}
\qquad [31]
$$

in which

$$\lambda = \frac{1}{\mathrm{H}} \qquad [32]$$

On the other hand, if Eqs. 30 are multiplied by $x_1, x_2, \cdots x_n$ respectively and added, Eq. 10 being noted, there results

$$(x_1^2 + x_2^2 + \cdots + x_n^2) - \mathrm{H}F = 0 \qquad [33]$$

and in view of the condition 27, this yields

$$\mathrm{H} = (x_1^2 + x_2^2 + \cdots + x_n^2) = |x|^2 \qquad [34]$$

Equations 31, 32, and 34 together represent the solution to the maximum-minimum problem. The results are seen to be extremely interesting, since Eqs. 31 are recognized as the conditions for which the direction of the vector x is left unchanged by the linear transformation 1, as discussed in Art. 10 and 12, Ch. III. The results of the present problem, therefore, yield an alternative geometrical interpretation for these directions, namely: *The principal axes of the quadric surface defined by Eqs. 2 and 15 together are those directions in space which are invariant to the symmetrical transformation 1.*

According to the discussion in Art. 10, Ch. III, the direction cosines of the principal axes are, therefore defined by Eq. 147 of that chapter, in which the $\overset{s}{k}_{ik}$ are the cofactors of the elements of the determinant in Eq. 146 for the latent roots λ_s. Each latent root defines a corresponding

principal axis, and the modal matrix $\mathfrak{L}$, given by Eq. 151, yields the directions of the complete set of n principal axes.

Since in the present discussions the matrix $\mathfrak{A}$ of the transformation 1 is symmetrical, the modal matrix $\mathfrak{L}$ is orthogonal, as shown in Art. 12, Ch. III. In other words, *the principal axes of the quadric surface forms a mutually orthogonal set.*

According to Eq. 32, there are n values for the Lagrangian multiplier

$$\mathrm{H}_s = \frac{1}{\lambda_s} \tag{35}$$

corresponding to the n latent roots of the matrix $\mathfrak{A}$. The quantities H_1, $\mathrm{H}_2, \cdots \mathrm{H}_n$, which are the reciprocals of the latent roots, are called the *proper values* (*eigenwerte*) of the matrix $\mathfrak{A}$ or of the corresponding quadratic form F. Since the latent roots for a symmetrical matrix are *real* (as shown in Art. 12, Ch. III), it is seen that the proper values of the quadratic form F are real also.

An interesting geometrical significance of these proper values is given by Eq. 34, which, for a particular vector $\overset{s}{x}$ appropriate to the root λ_s, reads

$$\mathrm{H}_s = (\overset{s}{x}_1{}^2 + \overset{s}{x}_2{}^2 + \cdots + \overset{s}{x}_n{}^2) = |\overset{s}{x}|^2 \tag{36}$$

and hence represents the square of the length of the corresponding principal axis extending from the origin to the quadric surface. These lengths are commonly referred to as the *semiaxes* of the quadric surface.

This result yields a geometrical interpretation to the latent roots of a symmetrical matrix. *They are the reciprocals of the squares of the semiaxes of the associated quadric surface.* If the latter is ellipsoidal, the lengths of all the semiaxes are real, and the squares of these lengths are positive. The quadric surface is, therefore, ellipsoidal when *all* the latent roots of the matrix $\mathfrak{A}$ are *positive*. The appearance of negative roots indicates that some of the lengths of the semiaxes are imaginary and hence that the surface has hyperbolic as well as ellipsoidal characteristics. If all the roots are negative, the surface is an imaginary ellipse, but it is customary to include this in the classification of completely ellipsoidal surfaces. In any case, the above maximum-minimum problem deals only in real values, since the square of the length of the vector x rather than its length enters into the manipulations.*

If the principal axes of the quadric surface are chosen as a new co-ordinate system, and if the variables $x'_1 \cdots x'_n$ refer to this system, the co-ordinate transformation from the original variables x_k to the new variables x'_k is given by the matrix equation

$$x] = \mathfrak{L}x'] \tag{37}$$

*The discussion in Art. 10 is also relevant to these matters.

According to Eqs. 16 and 25, the matrix of the quadratic form F for the new variables is

$$\mathfrak{L}_t \times \mathfrak{A} \times \mathfrak{L} \tag{38}$$

But by Eq. 208 of Art. 12, Ch. III, this is the diagonal matrix

$$\Lambda = \mathfrak{L}_t \times \mathfrak{A} \times \mathfrak{L} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \cdots & \lambda_n \end{bmatrix} \tag{39}$$

Hence the quadratic form F expressed in terms of the new variables x'_k reads

$$\begin{aligned} F &= \underline{x'}\Lambda x'] \\ &= \lambda_1 x'^2_1 + \lambda_2 x'^2_2 + \cdots + \lambda_n x'^2_n \\ &= \frac{x'^2_1}{\mathrm{H}_1} + \frac{x'^2_2}{\mathrm{H}_2} + \cdots + \frac{x'^2_n}{\mathrm{H}_n} \end{aligned} \tag{40}$$

Equated to unity, this is the equation of the quadric surface in the rectangular co-ordinate system which is coincident with the principal axes of that surface. It is referred to as the *normal form* for the equation of the surface. At the same time, Eq. 40 represents the reduction of the quadratic form F to a sum of squares.

This reduced form for F, together with the geometrical interpretation of the process of reduction, now yields a basis for interpreting the significance of the occurrence of zero roots or coincident roots to the characteristic equation of a *symmetrical* matrix. Thus coincident λ-roots evidently indicate a certain degeneracy of the associated quadric surface. For example, in three dimensions a coincidence of two of the latent roots in the case of an ellipsoid results in an ellipsoid of revolution, whereas for a coincidence of all three latent roots the ellipsoid of revolution degenerates into a sphere.

When two of the latent roots are coincident, it thus becomes clear from the discussion of simultaneous homogeneous equations in Art. 7, Ch. III, that the characteristic matrix $(\mathfrak{A} - \lambda\mathfrak{U})$, with λ equal to the value of the coincident root, must have the rank $n - 2$.

This reduction in rank of the characteristic determinant is required because a repeated nonzero λ-root would otherwise suggest that intrinsically less than n independent axes were required to describe the associated quadric surface. In other words, there would be less than n independent directions for which the vector describing the quadric surface goes through extrema. It is geometrically clear, however, that in the case of repeated nonzero λ-roots, the quadric surface still occupies intrinsically n dimensions in space, and therefore requires n independent

axes to describe it. Although the symmetry resulting from the equality of the *lengths* of two or more of the semiaxes of the surface means that the *directions* of these semiaxes are not unique, it must nevertheless be possible to pick correspondingly two or more independent (in fact, orthogonal) directions along which one may assign axes to describe the surface. Hence if there are p repeated nonzero λ-roots in a symmetric matrix describing a quadric surface in n dimensions, it must somehow be possible to find exactly p independent, but not necessarily unique, solutions for the $\overset{s}{x}$ corresponding to the repeated root. This fact requires that the characteristic matrix have rank $n - p$ for the value of λ in question. Under such circumstances the discussion in Art. 10, Ch. III, gives the procedure for finding such a set of n directions, and the corresponding nonsingular modal matrix $\mathfrak{L}$ which describes them. It is necessary to point out, however, that this geometrical discussion based upon quadratic forms is valid only for symmetric matrices, and, as pointed out in the article mentioned above, it is not possible to extrapolate the conclusions thus drawn regarding coincident λ-roots to the general case of nonsymmetric matrices.*

When the matrix $\mathcal{A}$ of the transformation 1 is singular — more specifically, if the rank of the matrix $\mathcal{A}$ is $n - p$ — it is possible to satisfy Eqs. 31 in p-independent ways for $\lambda = 0$, which means that p of the latent roots are zero, and the corresponding p proper values H_s (squares of the semi-axes of the quadric surface) are infinite. Equation 40 for the quadratic form F then has only $n - p$ terms, and the associated quadric surface is again degenerate, but in a way somewhat different from its degeneracy in the case of repeated nonzero roots. For example, a three-dimensional ellipsoid of rank 2 (the matrix $\mathcal{A}$ has the order 3 and the rank 2) is an elliptic cylinder because one of its semiaxes is infinite. It is still possible to find a nonsingular modal matrix $\mathfrak{L}$ which reduces the original matrix to diagonal form because of the fact that p independent vectors $\overset{s}{x}$ may still be found for $\lambda = 0$. The procedure is exactly that described in Art. 7, Ch. III, and corresponds to the method used for nonzero repeated λ-roots. Moreover, the converse of the present statement is also true; namely, that the occurrence of a zero λ-root of order p in a symmetric matrix of order n requires that the rank of the matrix be exactly $n - p$. The geometric reasoning substantiating this assertion is, again, that it must be possible to find n dimensions intrinsically occupied

* In general such a nonsymmetric matrix cannot be reduced to diagonal form by a collineatory transformation, since no nonsingular modal matrix may be found. An indication of what can be done by way of reduction may be found in G. Birkhoff and S. MacLane, *A Survey of Modern Algebra* (New York: The Macmillan Co., 1941), pp. 307–308, or R. Courant and D. Hilbert, *Methoden der mathematischen Physik* (Berlin: Julius Springer, 1931), Vol. I, Ch. I, pp. 36–37.

by the surface, even if p of them merely indicate co-ordinates upon which the function describing the surface is actually not dependent.

Thus it may be stated that *a given quadratic form of rank* r, *when reduced to a sum of squares, contains* r *terms.* That is, the new variables $x'_1 \cdots x'_r$ are only r in number whereas there are n original variables $x_1 \cdots x_n$. This result agrees with the geometrical interpretation because the description of an elliptic cylinder, in the normal form for example, requires only two variables inasmuch as the function describing the cylinder is independent of the longitudinal dimension.

5. A RELATED MAXIMUM-MINIMUM PROBLEM

A problem complementary to that treated in the preceding article is to determine the maxima or minima of the quadratic form F subject to the condition

$$\lfloor x\, x] = x_1^2 + x_2^2 + \cdots + x_n^2 = 1 \tag{41}$$

imposed upon its variables. Here the function

$$f(x_1, \cdots x_n) = F - \lambda(x_1^2 + x_2^2 + \cdots + x_n^2 - 1) \tag{42}$$

in which the Lagrangian multiplier is denoted by λ, represents the quadratic form provided the condition 41 is fulfilled.

The partial derivatives of $f(x_1, \cdots x_n)$ equated to zero are

$$\frac{\partial F}{\partial x_k} - 2\lambda x_k = 0 \quad \text{for } k = 1, 2, \cdots n \tag{43}$$

In view of Eq. 9, these again yield Eqs. 31. Multiplying Eqs. 43 respectively by $x_1, x_2, \cdots x_n$, adding them, and noting Eq. 10 gives

$$F - \lambda(x_1^2 + x_2^2 + \cdots + x_n^2) = 0 \tag{44}$$

Because of the condition 41, this is

$$F = \lambda \tag{45}$$

The conclusion is that the maxima or the minima of the quadratic form F, subject to the condition 41, occur for values of the variables defining the vectors $\overset{s}{x}$ whose directions are invariant to the linear transformation 1. Equation 45 shows that the corresponding maximum or minimum values of the quadratic form are equal to the latent roots of its matrix, that is,

$$F_{\substack{max\\min}} = \lambda_s \qquad (s = 1, 2, \cdots n) \tag{46}$$

In the present instance, the quadratic form F is not restricted to a constant and hence cannot be said to represent a quadric surface at all.

The tip of the vector y, on the other hand, does trace out a quadric surface as a result of Eq. 41 and the relation 18 between $x]$ and $y]$. The extrema of $|y|^2$, or the squares of the semiaxes of this surface, are given by the reciprocals of the latent roots of its matrix $\mathcal{A}^{-2}$ (Art. 4 of this chapter). But, by reason of the thoughts underlying the proof of the Cayley-Hamilton theorem (Art. 11, Ch. III), these roots are the reciprocals of the squares of the latent roots of $\mathcal{A}$. Hence the extrema of $|y|$, or the semiaxes of the associated quadric surface, are just the latent roots λ_s of Eq. 46. Thus the quadric surface associated with the transformation 1 may be alternatively thought of as the locus of the tip of the vector y, when the tip of the vector x is restricted to lie on the surface of an n-dimensional sphere, Eq. 41. The semiaxes of this surface may however be calculated in both magnitude and direction from the present conditions yielding the extrema of F subject to the condition 41. It is of interest to observe in this connection that the surface in question is always ellipsoidal, regardless of the nature of the nonsingular matrix $\mathcal{A}$.

6. An Interesting Application of These Results

In view of the results of this alternative maximum-minimum problem, a further particular question may be answered regarding the transformation 1. In addition to inquiring whether it is possible for the transform of x to be zero, or to have the same direction as x, one may ask: Is it possible for the transform of x to be a vector at right angles to x? In other words, under what conditions is the scalar product $x \cdot y$ zero when neither x nor y is zero?

It is shown by Eq. 14 that this scalar product is equal to the quadratic form F. Hence the present question amounts to inquiring whether F may vanish for a nonzero vector x, that is, for

$$x_1^2 + x_2^2 + \cdots + x_n^2 = k \neq 0 \qquad [47]$$

In the particular application to be considered, it will suffice to find the answer to this question for the special case in which the quadratic form F is positive definite (cannot go negative). Under such conditions, the places where F becomes zero are clearly minima of the quadratic form, and the question originally asked is equivalent to asking for the conditions under which a stationary point of F shall occur for $F = 0$.*

*In the case where F is not a positive definite form, it is not necessary for its zeros to occur at stationary points (subject to condition 47). The example $F = x_1^2 - x_2^2$ illustrates the point in question, when coupled with the condition $x_1^2 + x_2^2 = k \neq 0$. F is zero only when $x_1^2 = x_2^2 = \frac{k}{2}$, and the corresponding values of $y]$ are $y_1^2 = y_2^2 = \frac{k}{2}$. The stationary points of F are, however, at $x_2 = 0, x_1^2 = k$ and $x_1 = 0, x_2^2 = k$. Corresponding values of F and $y]$ are $y_2 = 0, y_1^2 = k$ and $y_1 = 0, y_2^2 = k$; while $F = k$ and $F = -k$, respectively.

If, then, condition 47 is substituted in place of Eq. 41, the above analysis remains essentially unaltered. In particular, Eq. 45 is replaced by

$$F = k\lambda \tag{48}$$

The answer is that F may be zero for $k \neq 0$ only when one or more of the latent roots are zero. The occurrence of a zero root means that the matrix $\mathcal{A}$ is singular. If at the same time Eqs. 1 are to have solutions for a nonzero y-vector, the latter must be a member of the transposed vector set of $\mathcal{A}$, which is the same as the vector set of $\mathcal{A}$ since $\mathcal{A}$ is symmetrical. Although the vector x constituting a solution is orthogonal to y, it is not simultaneously orthogonal to all the vectors in the set of $\mathcal{A}$, since y would then necessarily be zero.*

A useful application of these thoughts is to the problem of determining whether a given vector set is linearly dependent or independent. The m vectors $v_1, v_2, \cdots v_m$ in an n-dimensional space, expressed in terms of their components, yield the nonsquare matrix

$$\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix} \tag{49}$$

for which it is to be assumed that $m < n$.

To determine whether this vector set is linearly dependent or not requires, according to the discussion in Art. 2, Ch. III, that all m-square determinants be formed from the matrix 49 by selecting m columns in all possible combinations. The linear dependence of the vector set is established only if *all* these determinants are found to be zero.

A considerably less laborious procedure is the following. The vector set is linearly dependent if there exists a relation of the form

$$x_1v_1 + x_2v_2 + \cdots + x_mv_m = 0 \tag{50}$$

in which at least one of the coefficients x_k is not zero. The latter condition may be replaced by the requirement that

$$x_1^2 + x_2^2 + \cdots + x_m^2 = 1 \tag{51}$$

which is neither more nor less binding than to require that at least one of the x_k's be other than zero.

The quadratic form obtained by forming the scalar product of the vector expression in Eq. 50 with itself may be written

$$F = (x_1v_1 + x_2v_2 + \cdots + x_mv_m)^2 = \sum_{i,k=1}^{n} (v_i \cdot v_k)x_ix_k \geqq 0 \tag{52}$$

Fulfillment of Eq. 50 requires that the quadratic form F be zero. Hence

*These matters are discussed in Art. 7, Ch. III.

if the given vector set is linearly dependent, it must be possible to find $F = 0$ subject to the condition 51. In other words, one of the latent roots of F must be equal to zero. According to the previous discussion this fact requires that the discriminant of the form 52 be zero; hence that

$$G = \begin{vmatrix} v_1 \cdot v_1 & v_1 \cdot v_2 & \cdots & v_1 \cdot v_m \\ v_2 \cdot v_1 & v_2 \cdot v_2 & \cdots & v_2 \cdot v_m \\ \cdots & \cdots & \cdots & \cdots \\ v_m \cdot v_1 & v_m \cdot v_2 & \cdots & v_m \cdot v_m \end{vmatrix} = 0 \qquad [53]$$

Thus the vanishing of only a single m-square determinant (called the *Gramian* determinant) need be investigated in order to establish the dependence or independence of the given vector set.

7. Alternative reductions

From the discussion in Art. 4, it should be clear that the reduction of a quadratic form to a sum of squares is identical with the transformation of its square matrix to the diagonal form. The latter problem is discussed in Art. 10, Ch. II, where it is shown that when the matrix $\mathfrak{A}$ is symmetrical, reduction of it to a diagonal matrix $\mathfrak{D}$ is accomplished by the congruent transformation

$$\mathfrak{Q}_t \times \mathfrak{A} \times \mathfrak{Q} = \mathfrak{D} \qquad [54]$$

in which $\mathfrak{Q}$ is a nonsingular matrix, expressible as the product of elementary transformation matrices.

According to Eqs. 16 and 25, it is clear that the corresponding transformation from the variables $x_1 \cdots x_n$ to the new variables $x'_1 \cdots x'_n$ is given by

$$x] = \mathfrak{Q}x'] \qquad [55]$$

Hence if the diagonal matrix is written

$$\mathfrak{D} = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & & d_{nn} \end{bmatrix} \qquad [56]$$

then the quadratic form F, Eq. 2, expressed in terms of the new variables, reads

$$F = d_{11}x'^2_1 + d_{22}x'^2_2 + \cdots + d_{nn}x'^2_n \qquad [57]$$

More generally, if the rank of the matrix $\mathfrak{A}$ (rank of the quadratic form) is r, the diagonal matrix $\mathfrak{D}$, after suitable arrangement of its rows,

has the form

$$\mathcal{D} = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_{rr} & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{58}$$

and the reduced quadratic form has r terms; thus

$$F = d_{11}x'_1{}^2 + d_{22}x'_2{}^2 + \cdots + d_{rr}x'_r{}^2 \tag{59}$$

The matrix $\mathcal{A}$, however, possesses no *unique* diagonal form. According to the procedure given in Art. 10, Ch. II, there are any number of transformation matrices which can take the place of $\mathcal{Q}$ in Eq. 54 and also effect a reduction of $\mathcal{A}$ to a diagonal form. For example, another nonsingular matrix $\mathcal{P}$ may be found such that

$$\mathcal{P}_t \times \mathcal{A} \times \mathcal{P} = \mathcal{D}' \tag{60}$$

in which $\mathcal{D}'$ is also a diagonal matrix. The transformation of variables leading to Eq. 60 may be represented by a relation similar to Eq. 55,

$$x] = \mathcal{P}x''] \tag{61}$$

Correspondingly, there are any number of reduced forms for F, such as the one given by Eq. 59.

In this connection it is important to note, however, that if the rank of the matrix $\mathcal{A}$ is r, *all* the diagonal matrices to which it may be reduced by a transformation of the form given by Eqs. 54 or 60 have the same number of nonzero diagonal elements (namely r), because all the transformation matrices, $\mathcal{Q}$, $\mathcal{P}$, $\cdots$ etc., are nonsingular. Consequently, all various diagonal matrices are equivalent to $\mathcal{A}$, and equivalent matrices have the same rank (see Art. 9, Ch. II).

Not only do the diagonal matrices $\mathcal{D}$ and $\mathcal{D}'$ have the same number of nonzero elements, but they must have the same number of positive elements as well. This important property may be demonstrated by first assuming the contrary, and thereby arriving at a contradiction.

Since the variables $x']$ and $x'']$ are uniquely related to the original variables $x]$ by nonsingular transformations 55 and 61, respectively, the quadratic forms resulting from the diagonal matrices $\mathcal{D}$ and $\mathcal{D}'$ must be identically equal for all corresponding values of $x']$ and $x'']$. If the elements of $\mathcal{D}$ are now written d_i, and those of $\mathcal{D}'$ are written d'_i, the identity mentioned above becomes

$$d_1(x'_1)^2 + \cdots + d_\mu(x'_\mu)^2 + d_{\mu+1}(x'_{\mu+1})^2 + \cdots + d_r(x'_r)^2 \equiv$$
$$d'_1(x''_1)^2 + \cdots + d'_\nu(x''_\nu)^2 + d'_{\nu+1}(x''_{\nu+1})^2 + \cdots + d'_r(x''_r)^2 \tag{62}$$

Let the notation be chosen so that the first μ terms in $\mathfrak{D}$ and the first ν terms in $\mathfrak{D}'$ are the positive elements in these matrices. The assumption that one of these numbers is greater may be stated by requiring either $\mu > \nu$ or $\nu > \mu$. It is again merely a matter of notation to assume $\nu < \mu(\leqq r \leqq n)$. The following special choice of variables in Eq. 62 is now made:

$$\begin{aligned} x'_{\mu+1} = x'_{\mu+2} = \cdots = x'_n = 0 \\ x''_1 = x''_2 = \cdots = x''_\nu = 0 \end{aligned} \qquad [63]$$

In terms of the original variables $x]$ in Eqs. 55 and 61, the conditions 63 yield $\nu + (n - \mu) < n$ equations in n unknowns. It is then surely possible to find a solution for $x]$ in which not all the variables are zero (Art. 7, Ch. III). Moreover, the fact that $\mathfrak{P}$ and $\mathfrak{Q}$ are nonsingular means that values of $x''_{\nu+1} \cdots x''_n$ and $x'_1 \cdots x'_\mu$ exist such that neither all the x''_i nor all the x'_i are zero. These special values of $x']$ and $x'']$ are now substituted into Eq. 62, and the fact that $d_1, \cdots, d_\mu > 0$ while $d'_{\nu+1}, \cdots, d'_r < 0$ is emphasized by using absolute values of the latter. The identity 62 then becomes the equation

$$d_1(x'_1)^2 + \cdots + d_\mu(x'_\mu)^2 = -|d'_{\nu+1}|(x''_{\nu+1})^2 - \cdots - |d'_r|(x''_r)^2 \qquad [64]$$

The result 64 is clearly impossible, since the left side is definitely greater than zero and the right side cannot possibly be greater than zero. It is therefore necessary to conclude that $\nu = \mu$, since the assumption of an inequality leads to the contradiction in Eq. 64.

All the various reduced forms for F, like the one given by Eq. 59, therefore, have two things in common regardless of how this reduction is accomplished. The total number of terms is always equal to the rank r, and the numbers of positive and of negative coefficients are always alike. This result is known as the *law of inertia* of quadratic forms.

If the number of positive coefficients in a reduced form is denoted by P, and the number of negative ones by N, evidently

$$r = P + N \qquad [65]$$

The difference

$$s = P - N \qquad [66]$$

is called the ***signature*** of a quadratic form. Both the rank and the signature are thus seen to be invariant to the congruent transformation given by Eq. 25, in which $\mathfrak{C}$ is any real nonsingular matrix.

Although there are any number of matrices like $\mathfrak{Q}$, $\mathfrak{P}$, $\cdots$, etc., which reduce the matrix $\mathfrak{A}$ to a diagonal form and F to a sum of squares (according to the congruent transformation 54 or 60), there is essentially only one orthogonal matrix which effects such a reduction, namely the

modal matrix $\mathcal{L}$ defined in Art. 10, Ch. III.* Only for this orthogonal matrix do the coefficients of F in its reduced form equal the latent roots of $\mathcal{A}$. In other words, if the elements $d_{11}, d_{22}, \cdots d_{nn}$ in the reduced form given by Eq. 57 are equal to the latent roots $\lambda_1, \lambda_2, \cdots \lambda_n$ respectively, the matrix $\mathcal{Q}$ in Eq. 54 is the orthogonal modal matrix $\mathcal{L}$.

The congruent transformation, which applies to the transformation of the variables in a quadratic form (as discussed in Art. 3), should not be confused with the collineatory transformation discussed in Art. 10, Ch. III. Since the matrix $\mathcal{A}$ of a quadratic form is symmetrical, its modal matrix $\mathcal{L}$ is orthogonal (as shown in Art. 12, Ch. III). The collineatory transformation $\mathcal{L}^{-1}\mathcal{A}\mathcal{L}$, which carries $\mathcal{A}$ over into the diagonal form of its latent roots, is then identical with the congruent transformation $\mathcal{L}_t\mathcal{A}\mathcal{L}$, because the transpose and the inverse of an orthogonal matrix are the same.

However, if the columns of $\mathcal{L}$ are multiplied by arbitrary nonzero factors, the collineatory transformation still accomplishes the *same* result (as pointed out in Art. 10, Ch. III), but the matrix obtained by this modification of $\mathcal{L}$ is no longer orthogonal, and a congruent transformation of $\mathcal{A}$ by means of it does not reduce $\mathcal{A}$ to a diagonal form and hence does not reduce the corresponding quadratic form to a sum of squares.

Thus when a collineatory transformation reduces the symmetrical matrix $\mathcal{A}$ to the diagonal form of its latent roots, the matrix effecting this reduction is not necessarily an orthogonal one (as in the case of a congruent transformation), but may be an orthogonal matrix post-multiplied by an arbitrary nonsingular diagonal matrix.

8. Definite quadratic forms

If all the coefficients d_{kk} in the reduced form for F are positive or all negative, it is evident that the quadratic form is positive or negative respectively for all possible nonzero values of the variables. This result is true whether the variables are the x'_k appearing in the reduced form or the x_k related to the variables x'_k by a nonsingular transformation like Eq. 55, because this transformation amounts merely to a change in notation. Such quadratic forms are referred to as being either positive or negative *definite*.

If a quadratic form is positive definite, all the latent roots of its matrix are positive. Conversely, if all the latent roots are positive, the quadratic form must be positive definite. Incidentally, it is worth noting in this connection that the quadric surface of a positive definite form must be ellipsoidal.

*The only freedom in the construction of the matrix $\mathcal{L}$ occurs when there are repeated latent roots, and the directions of some of the principal axes of the corresponding quadratic form are not unique.

When the quadratic form is positive definite, so that all the coefficients d_{kk} in Eq. 59 are positive, the further congruent (real) transformation

$$x'_k = \frac{x''_k}{\sqrt{d_{kk}}} \qquad (\text{for } k = 1, 2, \cdots r) \tag{67}$$

evidently reduces F to the so-called *canonic* form

$$F = x''^2_1 + x''^2_2 + \cdots + x''^2_r \tag{68}$$

in which all the coefficients are $+1$. The requirement that F be positive definite is necessary in order that the transformation given by Eq. 67 be real. Hence the statement: *A positive definite quadratic form (singular or nonsingular) may always be reduced to the canonic form, given by Eq. 68, by means of a real nonsingular congruent transformation.*

9. A CRITERION FOR POSITIVE DEFINITENESS

If the quadratic form given by Eq. 2 is positive definite, it is readily seen that regardless of its rank, none of the diagonal elements of its matrix can be either negative or zero. For, if a_{kk} were such a zero or negative element, then setting all the variables except x_k and x_i equal to zero, and letting $x_i = 1$, would give the form the value

$$F = a_{ii} + 2a_{ik}x_k + a_{kk}x_k^2 \tag{69}$$

which, for $a_{kk} \leqq 0$, can certainly become negative by a suitable choice for the value of x_k. Although the condition

$$a_{kk} > 0 \tag{70}$$

is a necessary one for positive definiteness, it is not sufficient.

A useful set of necessary and sufficient conditions may be found through the following method of representing a positive definite quadratic form as a sum of squares. If the quadratic form is assumed to be nonsingular and positive definite, the canonic form reads

$$F = x'^2_1 + x'^2_2 + \cdots + x'^2_n \tag{71}$$

and the variables $x'_1 \cdots x'_n$ must be related to the variables $x_1 \cdots x_n$ by a linear transformation such as

$$\begin{aligned} p_{11}x_1 + p_{12}x_2 + \cdots + p_{1n}x_n &= x'_1 \\ p_{21}x_1 + p_{22}x_2 + \cdots + p_{2n}x_n &= x'_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots&\cdots\cdots \\ p_{n1}x_1 + p_{n2}x_2 + \cdots + p_{nn}x_n &= x'_n \end{aligned} \tag{72}$$

The matrix $\mathcal{P}$ of this transformation is restricted only by the condition that it be nonsingular and *real;* that is, the coefficients p_{ik} must be real

numbers. In matrix form, Eqs. 72 read

$$\mathcal{P}x] = x'] \tag{73}$$

and hence

$$F = \underline{x'}x'] = \underline{x}\mathcal{P}_t \times \mathcal{P}x] \tag{74}$$

Identifying this result with Eq. 21 for F, it follows that the matrix $\mathcal{A}$ is given by

$$\mathcal{A} = \mathcal{P}_t \times \mathcal{P} \tag{75}$$

If $p_1^t, p_2^t, \cdots p_n^t$ represent the transposed vector set of $\mathcal{P}$, Eq. 75 shows that the elements of the matrix $\mathcal{A}$ are given by the scalar products

$$a_{ik} = p_i^t \cdot p_k^t \tag{76}$$

It thus becomes clear that the elements of the matrix of a nonsingular positive definite quadratic form are necessarily determined from a linearly independent set of vectors in a manner similar to that which yields the elements of the Gramian determinant given by Eq. 53. The sufficiency of this conclusion is appreciated from the fact that if, for a given matrix $\mathcal{A}$, a nonsingular matrix $\mathcal{P}$ satisfying Eq. 75 can be found, the transformation of variables given by Eq. 73 reduces the equations having the matrix $\mathcal{A}$ (Eqs. 1) to a set of identities. That is, $\mathcal{P}_t^{-1}\mathcal{A}\mathcal{P}^{-1} = \mathcal{U}$ (the unit matrix of like order), and all the diagonal elements of $\mathcal{U}$ are $+1$. The existence of such a matrix $\mathcal{P}$ is thus recognized to be the necessary and sufficient condition for the positive definiteness of the quadratic form associated with the nonsingular matrix $\mathcal{A}$.

Incidentally, it may be noted that the elements of the matrix $\mathcal{G}$ of the fundamental metric tensor (see Art. 5, Ch. III), are formed in a similar manner. The latter is, therefore, the matrix of a positive definite quadratic form.

In particular, it may be observed that the elements on the principal diagonal of $\mathcal{A}$ are given by

$$a_{kk} = p_k^t \cdot p_k^t = |p_k^t|^2 \tag{77}$$

These are equal to the squares of the absolute lengths of the vectors. According to the remarks made in the opening paragraph of this article, none of the vectors p_k^t is, therefore, allowed to be zero. This, as well as the condition that the vector set p_k^t be linearly independent, is assured by the requirement that $\mathcal{P}$ be nonsingular.

It should now be observed that the assumption of a more special form for the matrix $\mathcal{P}$ does not subject the above argument to any restrictions. Thus the vector p_1^t in the transposed vector set of $\mathcal{P}$ may be assumed to be coincident with axis 1 of the rectangular reference system. The second

vector p_2' is chosen to lie in the plane determined by the axes 1 and 2; the third vector p_3' is then oriented so as to be confined within the three-dimensional subspace determined by the axes 1, 2, and 3, and so on. This procedure in no way restricts the generality of the vector set. $p_1' \cdots p_n'$, but merely amounts to a particular orientation of the reference axes relative to the given vector set p_k'.

The result of this choice of orientation for the reference axes is that p_1' has no components other than that on the axis 1; p_2' has no components other than those on the axes 1 and 2, and so forth. The matrix $\mathcal{P}$ then assumes the special triangular form

$$\mathcal{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ 0 & p_{22} & p_{23} & \cdots & p_{2n} \\ 0 & 0 & p_{33} & \cdots & p_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & p_{nn} \end{bmatrix} \qquad [78]$$

The real nonsingular matrix $\mathcal{P}$ having this form, the problem now is to discover the conditions which are imposed upon the given matrix $\mathcal{R}$ by the requirement that it shall have the representation expressed by Eq. 75 (according to which its elements are given by Eq. 76). The most direct way of discovering these conditions is to proceed with a determination of the matrix $\mathcal{P}$ for a given matrix $\mathcal{A}$. A possible procedure for this determination is suggested by the method of reducing an arbitrary symmetrical matrix to a diagonal form by means of the congruent transformation 54 (as discussed in Art. 10, Ch. II, and there expressed by Eq. 189). According to this method, the matrix $\mathcal{Q}$ effecting the reduction may have the same triangular form as $\mathcal{P}$ in Eq. 78 if (as is true in the present problem) none of the diagonal elements in the given matrix $\mathcal{A}$ are zero. The matrix $\mathcal{Q}$, moreover, has diagonal elements which are all $+1$.

With reference to Eqs. 54 and 75 one may write

$$\mathcal{A} = \mathcal{Q}_t^{-1}\mathcal{D}\mathcal{Q}^{-1} = \mathcal{P}_t\mathcal{P} \qquad [79]$$

Since the inverse of the matrix $\mathcal{Q}$ again has the same triangular form with diagonal elements which are all $+1$ (as may readily be seen from any method of matrix inversion, for example, the one discussed in Art. 6, Ch. II), the formation of the matrix $\mathcal{P}$ is made evident by Eq. 79. Thus, denoting by $\mathcal{D}^{1/2}$ the diagonal matrix whose diagonal elements are the square roots of the respective elements of $\mathcal{D}$, one has

$$\mathcal{P} = \mathcal{D}^{1/2}\mathcal{Q}^{-1} \qquad [80]$$

The significant part about this result is that the diagonal elements of $\mathcal{P}$ are those of $\mathcal{D}^{1/2}$. The diagonal elements of $\mathcal{D}$, therefore, are the squares of the diagonal elements $p_{11}, p_{22}, \cdots p_{nn}$ of $\mathcal{P}$. The positiveness of these

squared elements is thus seen to be the necessary and sufficient condition for the positiveness of the diagonal elements in $\mathcal{D}$, and hence for the positive definiteness of the quadratic form having the matrix $\mathcal{A}$. It remains to express this requirement in the form of conditions upon the elements of the matrix $\mathcal{A}$.

The determinant P of the triangular matrix in Eq. 78 is equal to the product $(p_{11}p_{22} \cdots p_{nn})$ of its diagonal elements, none of which, because $\mathcal{P}$ is nonsingular, is allowed to be zero. According to Eq. 75, the determinant A of the matrix $\mathcal{A}$ is

$$A = P^2 = (p_{11}p_{22} \cdots p_{nn})^2 \tag{81}$$

Now if the variable x_n is set equal to zero, x'_n (according to the transformation 73 and the special form of $\mathcal{P}$ given by Eq. 78) becomes zero also. The quadratic form F then appears as a function of the variables $x_1 \cdots x_{n-1}$ only, or in the reduced form as a function of the variables $x'_1 \cdots x'_{n-1}$ only. The elements in the last row and column of the matrix $\mathcal{A}$ or of the matrix $\mathcal{P}$ then have no further influence upon the values of F in Eq. 2, so that the latter may be regarded as a quadratic form in $n - 1$ variables with a matrix $\mathcal{A}_{n-1}$ obtained through striking out the last row and column in $\mathcal{A}$. Correspondingly, the last row and column in $\mathcal{P}$ may be struck out, and the remainder denoted by $\mathcal{P}_{n-1}$. If P_{n-1} is the determinant of $\mathcal{P}_{n-1}$ and A_{n-1} that of $\mathcal{A}_{n-1}$, the same reasoning as before shows that

$$A_{n-1} = P_{n-1}^2 = (p_{11}p_{22} \cdots p_{n-1})^2 \tag{82}$$

In a like manner, through also setting x_{n-1}, x_{n-2}, etc., equal to zero, it may be seen that the determinants A_{n-2}, A_{n-3}, etc., which are obtained by striking out the last two, three, etc., rows and columns in A, are given by

$$\begin{aligned} A_{n-2} &= P_{n-2}^2 = (p_{11}p_{22} \cdots p_{n-2})^2 \\ A_{n-3} &= P_{n-3}^2 = (p_{11}p_{22} \cdots p_{n-3})^2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ A_1 &= a_{11} = p_{11}^2 \end{aligned} \tag{83}$$

From these equations it follows that

$$\begin{aligned} p_{11}^2 &= A_1 = a_{11} \\ p_{22}^2 &= \frac{A_2}{A_1} \\ p_{33}^2 &= \frac{A_3}{A_2} \\ &\cdots\cdots\cdots \\ p_{nn}^2 &= \frac{A}{A_{n-1}} \end{aligned} \tag{84}$$

The necessary and sufficient conditions that a nonsingular quadratic form F with the discriminant A be positive definite are, therefore, stated by the inequalities

$$A > 0, A_{n-1} > 0, A_{n-2} > 0, \cdots A_1 = a_{11} > 0 \qquad [85]$$

A nonsingular quadratic form is positive definite if the discriminant and all its principal minors have positive nonzero values.

The relations 84 together with Eq. 76 afford a method for determining the triangular matrix of Eq. 78 whereby the quadratic form may be reduced to the canonic form given by Eq. 71. Thus the diagonal elements in $\mathcal{P}$ are given by Eqs. 84 directly. Next, Eq. 76 shows that

$$a_{1k} = p_{11}p_{1k} \quad \text{for } k = 2, 3, \cdots n \qquad [86]$$

from which the remaining elements in the first row of $\mathcal{P}$ are determined.

Using Eq. 76 again, one finds

$$a_{2k} = p_{12}p_{1k} + p_{22}p_{2k} \quad \text{for } k = 3, 4, \cdots n \qquad [87]$$

Here p_{2k} is the only unknown, and so the remaining elements of the second row of $\mathcal{P}$ can be calculated. From

$$a_{3k} = p_{13}p_{1k} + p_{23}p_{2k} + p_{33}p_{3k} \quad \text{for } k = 4, 5, \cdots n \qquad [88]$$

in which p_{3k} is the only unknown, the remaining elements in the third row of $\mathcal{P}$ are found, and so forth.

If the quadratic form F is singular — more specifically, if its matrix $\mathcal{A}$ has the rank $n - p$ — then the criteria for positive definiteness are the same as those given by the relations 85 except that the equals sign is included with the first p inequalities.

Thus if $\mathcal{A}$ is the matrix of a positive definite quadratic form of rank r, it may be reduced to its canonic form by means of a real nonsingular triangular matrix $\mathcal{Q}$ in the congruent transformation

$$\mathcal{Q}_t\mathcal{A}\mathcal{Q} = \mathfrak{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & 0 \cdots & 0 \\ 0 & 0 & & & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & & \cdots & 0 \end{bmatrix} \qquad [89]$$

Here the canonic matrix $\mathfrak{S}$ has r units $(+1)$ on its principal diagonal. The matrix $\mathcal{Q}$ has the same triangular form as $\mathcal{P}$ in Eq. 78.

The inverse of $\mathcal{Q}$ again has the same triangular form, and hence may be

identified with $\mathcal{P}$ of Eq. 78, so that the singular matrix $\mathcal{a}$ is represented by

$$\mathcal{a} = \mathcal{P}_t \mathfrak{S} \mathcal{P} \tag{90}$$

in place of Eq. 75, which holds only when $\mathcal{a}$ is nonsingular.

Bearing in mind the form of the canonic matrix $\mathfrak{S}$ as shown in Eq. 89, and applying the same reasoning as in the previous argument in which $\mathcal{a}$ is assumed to have the rank n, one may establish the above statement regarding the modification of the criteria 85 for any rank $n - p$.

10. The iterated quadratic form

The quadratic form which results when the matrix $\mathcal{a}$ of F is replaced by the kth power of $\mathcal{a}$ is referred to as the *iterated* form of kth order, or as the kth iterated form of F. If $\mathfrak{L}$ is the orthogonal modal matrix of $\mathcal{a}$, as pointed out in Art. 11, Ch. III,

$$\mathfrak{L}_t \mathcal{a}^2 \mathfrak{L} = \mathfrak{L}_t \mathcal{a} \mathfrak{L} \mathfrak{L}_t \mathcal{a} \mathfrak{L} = \Lambda^2 \tag{91}$$

in which Λ is the diagonal matrix of the latent roots of $\mathcal{a}$.

Since

$$\Lambda^2 = \begin{bmatrix} \lambda_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & & \lambda_n^2 \end{bmatrix} \tag{92}$$

the latent roots of the second iterated form of F are equal to the squares of the latent roots of F. More generally, the latent roots of the kth iterated form of F are recognized as being equal to the kth powers of the latent roots of F.

It is also clear that the modal matrix $\mathfrak{L}$ which transforms F to its normal form (sum of squares) also transforms any of the iterated forms of F to their normal forms, and that the latter differ from the normal form of F only in that the coefficients of the square terms appear raised to the kth power.

It appears that all the iterated forms of *even* order are *positive definite*, and that their associated quadric surfaces are entirely ellipsoidal in character. Moreover, the principal axes of these ellipsoidal quadrics coincide in direction with those of the surface associated with F, since the modal matrix $\mathfrak{L}$ defines these directions.

Hence in searching for the principal axes of a quadric surface $F = 1$, it is immaterial whether the given form F or any of its iterated forms is considered. With regard to the method discussed in Art. 4, it is thus recognized that the argument is not subjected to any loss in generality by the assumption, for the purpose of clarifying the geometrical visualization, that the given quadric surface is entirely ellipsoidal in character.

11. The simultaneous reduction of a pair of quadratic forms to sums of squares

Two quadratic forms are considered to be given by the matrix equations

$$F_1 = \underbrace{x} \mathcal{A} x] \tag{93}$$

and

$$F_2 = \underbrace{x} \mathcal{B} x] \tag{94}$$

In the present discussion, it is assumed that at least one of these forms is positive definite and nonsingular.*

The fundamental principle upon which the simultaneous reduction is based is the following. If one of the forms, for example F_2, is nonsingular and positive definite, the associated quadric surface may be visualized as that of an ellipsoid. The variables $x_1 \cdots x_n$ of the two functions F_1 and F_2 are first subjected to the orthogonal transformation which introduces the principal axes of this ellipsoid as a new co-ordinate system. This transformation reduces F_2 to a sum of squares but does not in general eliminate the cross-product terms from F_1 also.

A further real transformation of the form given by Eq. 67 next carries F_2 over into its canonic form, and the associated ellipsoid becomes a sphere. If F_2 were not nonsingular, the resulting quadric surface would not be spherical; and if F_2 were not positive definite, the transformation which reduces it to its canonic form would not be real. The resulting form F_1 at this stage would then not be real either, and the following step could not be carried out by means of a real transformation.

The equation of a sphere with its center at the origin is evidently in the normal form (sum of squares with coefficients unity) regardless of the angular orientation of the rectangular co-ordinate axes. Any set of orthogonal axes can be principal axes for the sphere. Hence the variables in both quadratic forms can now be subjected to any orthogonal transformation without further affecting the form of F_2. At this step an orthogonal transformation can be found which reduces F_1 to a sum of squares (transforms to the co-ordinates which are the principal axes of the quadric surface defined by the form which F_1 has at this stage), and F_2 remains in its canonic form.

A transformation which combines these three steps evidently accomplishes the desired simultaneous reduction of both F_1 and F_2. This

*In physical problems, quadratic forms usually represent energy functions, and hence this assumption does not constitute a serious restriction from the practical standpoint. The algebraically more complicated problem which results when these restrictions are not imposed, is not treated here [for an introduction to this more general problem the reader is referred to M. Bocher, *Introduction to Higher Algebra* (New York: The Macmillan Co., 1927)].

reduction is now formulated explicitly by means of the following matrix equations.

If F_2 is positive definite and nonsingular, the discussion in Art. 9 shows that its matrix $\mathcal{B}$ may be represented by the relation

$$\mathcal{B} = \mathcal{P}_t\mathcal{P} \tag{95}$$

in which $\mathcal{P}$ is a nonsingular real matrix. This matrix may be considered to be any one which satisfies the congruent transformation

$$\mathcal{P}_t^{-1}\mathcal{B}\mathcal{P}^{-1} = \mathcal{U} \tag{96}$$

$\mathcal{U}$ being the unit matrix having the same order as $\mathcal{B}$.

One method of finding $\mathcal{P}$ is to determine first the orthogonal modal matrix $\mathfrak{L}^{(b)}$ of $\mathcal{B}$ which, by means of the congruent transformation

$$\mathfrak{L}_t^{(b)}\mathcal{B}\mathfrak{L}^{(b)} = \Lambda_b \tag{97}$$

reduces $\mathcal{B}$ to the diagonal form of its latent roots. This diagonal matrix is denoted by Λ_b. Now if $\Lambda_b^{1/2}$ represents a diagonal matrix with diagonal elements equal to the square roots of the latent roots of $\mathcal{B}$ (none of which is zero),

$$\Lambda_b^{-1/2}\mathfrak{L}_t^{(b)}\mathcal{B}\mathfrak{L}^{(b)}\Lambda_b^{-1/2} = \mathcal{U} \tag{98}$$

A comparison of Eqs. 96 and 98 yields

$$\mathcal{P} = \Lambda_b^{1/2}\mathfrak{L}_t^{(b)} \tag{99}$$

However, since

$$\mathcal{O}_t\mathcal{U}\mathcal{O} = \mathcal{U} \tag{100}$$

in which $\mathcal{O}$ is an arbitrary orthogonal matrix of the same order as $\mathcal{U}$ and $\mathcal{B}$, it is readily seen that $\mathcal{P}$ may more generally be represented as

$$\mathcal{P} = \mathcal{O}_t\Lambda_b^{1/2}\mathfrak{L}_t^{(b)} \tag{101}$$

The matrix $\mathcal{P}$ may be found by this method, or perhaps more simply by means of the process discussed in Art. 10, Ch. II. This part of the simultaneous reduction of F_1 and F_2 is not unique and may be accomplished by whatever procedure appears to be the most expedient under the given circumstances.

If the variables $x_1 \cdots x_n$ in the quadratic forms given by Eqs. 93 and 94 are subjected to the transformation expressed by the matrix equation

$$x] = \mathcal{P}^{-1}x'] \tag{102}$$

the matrices of these quadratic forms are subjected to the congruent transformations

$$\mathcal{P}_t^{-1}\mathcal{A}\mathcal{P}^{-1} = \mathcal{A}^{(b)} \tag{103}$$

and

$$\mathcal{P}_t^{-1}\mathcal{B}\mathcal{P}^{-1} = \mathcal{U} \qquad [104]$$

F_2 is thus reduced to its canonic form. F_1 at this stage has the matrix $\mathcal{A}^{(b)}$, which evidently is still real and symmetrical. Hence it possesses an orthogonal modal matrix $\mathcal{L}^{(a\ b)}$ which carries it over into the diagonal form of its latent roots; thus

$$\mathcal{L}_t^{(a\ b)}\mathcal{A}^{(b)}\mathcal{L}^{(a\ b)} = \Lambda_{a\ b} \qquad [105]$$

These roots are called the latent roots of $\mathcal{A}$ *with respect to* $\mathcal{B}$. Since $\mathcal{A}^{(b)}$ is symmetrical, these latent roots are *real*. They are positive if $\mathcal{A}$ (as well as $\mathcal{B}$) is the matrix of a positive definite quadratic form (see Art. 7 of this chapter).*

The further transformation of variables

$$x'] = \mathcal{L}^{(a\ b)}x''] \qquad [106]$$

with the orthogonal matrix $\mathcal{L}^{(a\ b)}$, leaves F_2 in the canonic form

$$F_2 = (x''_1)^2 + (x''_2)^2 + \cdots + (x''_n)^2 \qquad [107]$$

and reduces F_1 to the sum of squares

$$F_1 = \lambda_1^{(a\ b)}(x''_1)^2 + \lambda_2^{(a\ b)}(x''_2)^2 + \cdots + \lambda_n^{(a/b)}(x''_n)^2 \qquad [108]$$

in which $\lambda_1^{(a,b)} \cdots \lambda_n^{(a\ b)}$ are the latent roots of $\mathcal{A}$ with respect to $\mathcal{B}$.

These are the roots of the characteristic equation

$$|\mathcal{A}^{(b)} - \lambda\mathcal{U}| = |\mathcal{P}_t^{-1}\mathcal{A}\mathcal{P}^{-1} - \lambda\mathcal{U}| = 0 \qquad [109]$$

Since the latent roots of a matrix are invariant to a collineatory transformation (see Art. 10, Ch. III), and

$$\mathcal{P}_t\mathcal{A}^{(b)}\mathcal{P}_t^{-1} = \mathcal{A}\mathcal{P}^{-1}\mathcal{P}_t^{-1} = \mathcal{A}\mathcal{B}^{-1} \qquad [110]$$

or

$$\mathcal{P}^{-1}\mathcal{A}^{(b)}\mathcal{P} = \mathcal{P}^{-1}\mathcal{P}_t^{-1}\mathcal{A} = \mathcal{B}^{-1}\mathcal{A} \qquad [111]$$

it follows that the characteristic equation 109 may alternatively be written in either of the forms

$$|\mathcal{A}\mathcal{B}^{-1} - \lambda\mathcal{U}| = 0 \qquad [112]$$

or

$$|\mathcal{B}^{-1}\mathcal{A} - \lambda\mathcal{U}| = 0 \qquad [113]$$

Since $\mathcal{B}$ is assumed to be nonsingular, the equation

$$|\mathcal{B}(\mathcal{B}^{-1}\mathcal{A} - \lambda\mathcal{U})| = 0 \qquad [114]$$

*This fact is an important one in the study of oscillating systems in dynamics and electrical network theory.

has the same roots as Eq. 113. Hence the latent roots of $\mathcal{A}$ with respect to $\mathcal{B}$ may also be defined as the roots of

$$|\mathcal{A} - \lambda\mathcal{B}| = 0 \qquad [115]$$

In other words, the characteristic equation for these latent roots may be written in any one of the four forms given by Eqs. 109, 112, 113, or 115.

Since there are no restrictions on the symmetrical matrix $\mathcal{A}$ (except that it be real, of course), some of the roots of the characteristic equation 115 may be negative, or coincident, or zero. Observe that $\mathcal{A}$ and $\mathcal{A}^{(b)}$ contained in Eq. 103 are equivalent matrices (because $\mathcal{P}$ is nonsingular) and hence they have the same rank. Therefore, if the rank of the matrix $\mathcal{A}$ of F_1 is r, the characteristic equation 109 or 115 has r nonzero roots, and the reduced form for F_1 given by Eq. 108 has r terms. The canonic form for F_2, however, must have n terms because the matrix of F_2 is assumed to be nonsingular.

12. An alternative geometrical interpretation of the same problem

So far in the discussions of this chapter, the various geometrical interpretations have assumed a rectangular co-ordinate system. Thus the equation $F = 1$ for the quadric surface associated with a given quadratic form is visualized geometrically by supposing the variables $x_1 \cdots x_n$ to be the co-ordinates of a point with respect to a rectangular system of axes. The normal form for this equation, for which F appears as a sum of squares with arbitrary real coefficients, is then familiarly recognized as yielding the quadric surface (ellipsoid, for example) with its orthogonal set of principal axes coincident with the co-ordinate axes.

Although the tacit assumption of a rectangular co-ordinate system is quite in order, it is just as feasible to suppose that the reference co-ordinate system, which is used for the geometrical interpretations, is given by an oblique set of axes, but these interpretations must then be revised.

First, in this regard, it is significant to observe that an ellipse, for example, remains an ellipse when the angle between the co-ordinate axes is allowed to depart from 90 degrees. The equation of the ellipse is supposed to remain fixed, and the co-ordinates x_1 and x_2 are then assumed to be the parallel projections of a point P upon the oblique axes; that is, they are the contravariant components of the vector OP (O denotes the origin) which is the vector sum of its components according to the familiar parallelogram law of addition. If the given equation is plotted by laying off corresponding values of x_1 and x_2 along such a set of axes, the geometrical form of the resulting figure is still elliptic provided it is elliptic

when identically the same equation is plotted by laying off x_1 and x_2 along a set of rectangular axes.

Next, it should be noted that even if the equation of the central ellipse is in its normal form, the principal axes of the ellipse need not coincide with any of the oblique co-ordinate axes. This fact the reader may readily establish for himself by plotting a simple numerical example on oblique axes. In this connection it should be observed that the equation of a circle in rectangular co-ordinates represents an ellipse when plotted in oblique co-ordinates.

According to the previous discussions of this chapter, the equation of a given central ellipse is transformed to its normal form by rotation of the rectangular co-ordinate system until it coincides with the principal axes of that ellipse. In view of the remarks of the present article, it appears that this reduction may alternatively be thought of as accomplished through transforming to an oblique set of co-ordinate axes whose angular orientations relative to the given ellipse are such as to yield this same ellipse by means of an equation which contains square terms only.

It is significant to note that there are an infinite number of pairs of oblique axes for which the equation of an ellipse goes into the normal form. The distinguishing characteristic of a normal form is the property that a reversal of the sign of any particular co-ordinate of a point on the curve always yields another point on the curve. If any line through the center of the figure is chosen as one of the oblique axes, the other axis required to yield an equation for the ellipse in normal form is uniquely determined. The symmetry requirement mentioned above may be met if the second axis is chosen to be the one bisecting all chords of the figure parallel to the first axis. Such a pair of axes is known as a pair of conjugate diameters, and it is a property of these lines that each bisects all chords parallel to the other. It is this property which meets the symmetry condition characterizing a normal form in an oblique set of co-ordinates. A simple way to find the second axis, when any first axis is arbitrarily assigned, follows from the realization that the tangent to the ellipse at either end of the first axis is a limiting form for the set of parallel chords bisected by that axis. Hence the second axis is assigned to be that line through the center of the ellipse which is parallel to a tangent to the curve at the extremity of the first diameter chosen.

The clue to the simultaneous reduction of two quadratic forms to sums of squares also lies in these ideas. Thus, if two ellipses are given (the two forms F_1 and F_2 may for the moment be thought of as defining a pair of central two-dimensional ellipses) with arbitrary semiaxes and with an arbitrary relative orientation, it is possible to find a single pair of oblique axes such that both of the given ellipses may be described by means of equations which contain square terms only.

A sketch of any two such central ellipses will convince the reader that it is always possible to choose a first axis in such a way that tangents to both ellipses, at the points where this line intersects them, are parallel. The second axis is then taken parallel to these tangents, and in this set of oblique co-ordinates both ellipses have the symmetry property which indicates a normal form for the equations representing them.

The feasibility of obtaining such a set of oblique axes in the general case is thus indicated, though the most expedient method of handling the problem analytically does not directly parallel the above geometrical discussion.

With reference to the Eqs. 93 and 94 for the forms F_1 and F_2, these ideas are formulated more specifically by the statement that there exists an oblique co-ordinate system defined by the unit vector set $c_1, c_2, \cdots c_n$ of the matrix $\mathcal{C}$, such that a transformation to the contravariant variables of this system (see Eq. 43, Art. 5, Ch. III) by means of the equation

$$x] = \mathcal{C}_t\xi] \tag{116}$$

and its transpose

$$\underline{x} = \underline{\xi}\,\mathcal{C} \tag{117}$$

simultaneously carries the expressions for these two quadratic forms over into

$$F_1 = \underline{\xi}\,\mathcal{C}\mathcal{A}\mathcal{C}_t\xi] = \underline{\xi}\,\mathcal{D}_1\xi] \tag{118}$$

and

$$F_2 = \underline{\xi}\,\mathcal{C}\mathcal{B}\mathcal{C}_t\xi] = \underline{\xi}\,\mathcal{D}_2\xi] \tag{119}$$

in which $\mathcal{D}_1$ and $\mathcal{D}_2$ are diagonal matrices.

The orientations of these oblique axes, defined by the vector set of the matrix $\mathcal{C}$, are found in the solution to the following closely related problem. The linear transformations associated with the quadratic forms F_1 and F_2 may be written

$$\mathcal{A}x] = y] \tag{120}$$

and

$$\mathcal{B}x] = z] \tag{121}$$

The first of these equations transforms a vector x into a vector y, and the second transforms the same vector x into a vector z. The question may be raised whether there are any particular directions for the vector x (relative to a fundamental rectangular co-ordinate system) for which its two transforms y and z are coincident in direction (but not necessarily

coincident with x). This condition is expressed by the matrix equation

$$y] = \lambda z] \tag{122}$$

in which λ is a constant multiplier.

Substituting Eqs. 120 and 121 into Eq. 122 yields

$$(\mathfrak{A} - \lambda\mathfrak{B})x] = 0 \tag{123}$$

The question, therefore, amounts to inquiring whether nontrivial solutions may be found to the set of homogeneous equations

$$\left.\begin{array}{l}(a_{11} - \lambda b_{11})x_1 + (a_{12} - \lambda b_{12})x_2 + \cdots + (a_{1n} - \lambda b_{1n})x_n = 0 \\ (a_{21} - \lambda b_{21})x_1 + (a_{22} - \lambda b_{22})x_2 + \cdots + (a_{2n} - \lambda b_{2n})x_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ (a_{n1} - \lambda b_{n1})x_1 + (a_{n2} - \lambda b_{n2})x_2 + \cdots + (a_{nn} - \lambda b_{nn})x_n = 0\end{array}\right\} \tag{124}$$

The condition for the existence of such solutions is that the determinant of this system of equations shall vanish, that is,

$$|\mathfrak{A} - \lambda\mathfrak{B}| = 0 \tag{125}$$

This is identical with the characteristic equation 115 which determines the latent roots of $\mathfrak{A}$ with respect to $\mathfrak{B}$. If the determinant of Eqs. 124 for a particular root $\lambda_s^{(a/b)}$ is written

$$|\mathfrak{A} - \lambda\mathfrak{B}|^{(a/b)}_{\lambda=\lambda_s} \tag{126}$$

and the cofactors of the ith row of this determinant are denoted by $\overset{s}{k}_{ik}$, the direction cosines of a vector $\overset{s}{x}$ constituting a solution appropriate to the root $\lambda = \lambda_s^{(a/b)}$ are given by

$$l_{ks} = \frac{\overset{s}{k}_{ik}}{\sqrt{(\overset{s}{k}_{i1})^2 + (\overset{s}{k}_{i2})^2 + \cdots + (\overset{s}{k}_{in})^2}} \quad \text{for } k = 1, 2, \cdots n \tag{127}$$

provided all the λ-roots of Eq. 125 are distinct.

For such a vector $\overset{s}{x}$ with the components $\overset{s}{x}_1, \overset{s}{x}_2, \cdots \overset{s}{x}_n$ relative to the fundamental rectangular co-ordinate system, Eq. 123 may be written

$$\mathfrak{A}\overset{s}{x}] = \lambda_s^{(a/b)}\mathfrak{B}\overset{s}{x}] \tag{128}$$

The n equations of this form, for the n particular vectors $\overset{1}{x}, \overset{2}{x}, \cdots \overset{n}{x}$ appropriate to the n roots of Eq. 125, may be combined into one matrix equation by definition of the modal matrix

$$\mathfrak{L} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \tag{129}$$

the elements of which are the direction cosines defined by Eq. 127. Then

the complete set of equations like the one given by Eq. 128 is contained in

$$\mathcal{A}\mathcal{L} = \mathcal{B}\mathcal{L}\Lambda_{a/b} \tag{130}$$

in which $\Lambda_{a/b}$ is the diagonal matrix of the latent roots of $\mathcal{A}$ with respect to $\mathcal{B}$, that is, the same diagonal matrix as that appearing in Eq. 105.

If Eq. 130 is premultiplied on both sides by the transpose of $\mathcal{L}$, it reads

$$\mathcal{L}_t\mathcal{A}\mathcal{L} = (\mathcal{L}_t\mathcal{B}\mathcal{L})\Lambda_{a/b} \tag{131}$$

The resultant matrices $\mathcal{L}_t\mathcal{A}\mathcal{L}$ and $\mathcal{L}_t\mathcal{B}\mathcal{L}$ are necessarily symmetrical because $\mathcal{A}$ and $\mathcal{B}$ are symmetrical. Equation 131 implies that after the columns of the symmetrical matrix $\mathcal{L}_t\mathcal{B}\mathcal{L}$ are multiplied by a set of factors $\lambda_1^{(a/b)} \cdots \lambda_n^{(a/b)}$ respectively, the resulting matrix is still symmetrical. This condition can be possible only if $\mathcal{L}_t\mathcal{A}\mathcal{L}$ and $\mathcal{L}_t\mathcal{B}\mathcal{L}$ are both diagonal matrices. Hence if these are written

$$\mathcal{L}_t\mathcal{A}\mathcal{L} = \mathcal{D}_1 \tag{132}$$

and

$$\mathcal{L}_t\mathcal{B}\mathcal{L} = \mathcal{D}_2 \tag{133}$$

then a comparison with Eqs. 118 and 119 shows that the matrix $\mathcal{C}$ has been found, namely,

$$\mathcal{C} = \mathcal{L}_t \tag{134}$$

The direction cosines of the set of oblique co-ordinate axes, in terms of which the equations of the two quadric surfaces associated with the forms F_1 and F_2 appear as sums of squares, are given by Eq. 127. More specifically, the direction cosines of the sth oblique axis are the elements in the sth column of the matrix $\mathcal{L}$ defined by Eqs. 127 and 129. The set of unit vectors characterizing the oblique axes with reference to the fundamental rectangular co-ordinate system is the transposed vector set of $\mathcal{L}$.

Equations 131, 132, and 133 show that

$$\Lambda_{a/b} = \mathcal{D}_2^{-1}\mathcal{D}_1 \tag{135}$$

or, if the diagonal elements in $\mathcal{D}_1$ and $\mathcal{D}_2$ are denoted by $d_{ss}^{(1)}$ and $d_{ss}^{(2)}$ respectively, that

$$\lambda_s^{(a/b)} = \frac{d_{ss}^{(1)}}{d_{ss}^{(2)}} \tag{136}$$

The quadratic forms F_1 and F_2 given by Eqs. 93 and 94 are, therefore, simultaneously reduced to

$$F_1 = d_{11}^{(1)}\xi_1^2 + d_{22}^{(1)}\xi_2^2 + \cdots + d_{nn}^{(1)}\xi_n^2 \tag{137}$$

and

$$F_2 = d_{11}^{(2)}\xi_1^2 + d_{22}^{(2)}\xi_2^2 + \cdots + d_{nn}^{(2)}\xi_n^2 \tag{138}$$

when the variables $x_1 \cdots x_n$ are subjected to the linear transformation

$$x] = \mathfrak{L}\xi] \tag{139}$$

The matrix $\mathfrak{L}$ is, of course, not orthogonal, as is clear from the fact that it defines a set of oblique axes or as is alternatively made evident from Eq. 130, which can be written

$$\mathfrak{L}^{-1}(\mathfrak{B}^{-1}\mathfrak{A})\mathfrak{L} = \Lambda_{a/b} \tag{140}$$

It thus appears that $\mathfrak{L}$ is the modal matrix which reduces the dissymmetrical matrix $\mathfrak{B}^{-1}\mathfrak{A}$ to the diagonal form of its latent roots.

As shown in the previous article, these latent roots are real, but if $\mathfrak{A}$ has the rank r, then $n - r$ of them are zero. Since $\mathfrak{B}$ is nonsingular, none of the diagonal elements $d_{ss}^{(2)}$ is zero, but Eq. 136 shows that as many elements of the diagonal matrix $\mathfrak{D}_1$, are zero as there are zero λ-roots in the characteristic equation 125. Thus the reduced form 137 has r terms, and that given by Eq. 138 has n terms.

Since F_2 is positive definite, all the coefficients in Eq. 138 are positive. A further real transformation of the variables then carries F_2 and F_1 over into the forms given by Eqs. 107 and 108 respectively.

13. A FEW REMARKS REGARDING THE SIMULTANEOUS REDUCTION OF MORE THAN TWO QUADRATIC FORMS TO SUMS OF SQUARES

In the solution of many practical problems, it would be very desirable to be able to reduce more than two quadratic forms to sums of squares simultaneously. To do so is in general not possible, however, as the reader can readily appreciate by considering an attempt to extend the reasoning in the opening paragraphs of Art. 11 or 12 to more than two forms.

If p quadratic forms $F_1, F_2, \cdots F_p$ are given, it is, however, possible to reduce them simultaneously in the special case for which there exist $p - 2$ independent relations of linear dependence of the form

$$\gamma_1 F_1 + \gamma_2 F_2 + \cdots + \gamma_p F_p = 0 \tag{141}$$

in which at least one of the coefficients has a nonzero value. The condition 141 must, of course, hold for all values of the common variables $x_1 \cdots x_n$.

Cases of this sort which do occur in practice are the problems of reducing the three forms F_1, F_2, and F_3, when

$$F_1 = \gamma F_2 \tag{142}$$

or of reducing the four forms, F_1, F_2, F_3, F_4, when

$$F_1 = \gamma F_2 \qquad [143]$$

and

$$F_3 = \delta F_4$$

Evidently, a transformation which carries F_1 and F_3 over into their reduced forms simultaneously yields a similar reduction for F_2 and F_4.

14. The abridgment of a quadratic form that results from imposing linear constraints upon its variables

Consider for the moment a quadratic form F in only three dimensions, and visualize its associated quadric surface defined by $F = 1$ as a central ellipsoid. If one demands that one of the three variables x_1, x_2, x_2 be zero (for example, if one arbitrarily sets $x_3 = 0$), it is clear geometrically that the ellipsoid thereby is reduced to the two-dimensional ellipse given by the intersection of the ellipsoid with the co-ordinate plane normal to axis 3, i.e., the 1–2 plane. This ellipse is defined by an equation $\bar{F} = 1$ in which $\bar{F}$ is obtained from F through simply dropping those terms involving x_3.

The equation $x_3 = 0$ may be regarded as a linear constraint which restricts the original variables to values corresponding to points on the intersection of the ellipse with the 1–2 plane, the latter being referred to as the *constraint plane*. $\bar{F}$, which is a quadratic form in only two variables, is spoken of as an *abridged* form of F.

The abridged form is not always so easily found as the very simple situation just described. Thus, with reference to the same three-dimensional form F, suppose that the constraint plane is chosen as an arbitrary one, still passing through the origin of co-ordinates, however. In this case the equation of constraint is changed from the simple form $x_3 = 0$ to

$$p_{31}x_1 + p_{32}x_2 + p_{33}x_3 = 0 \qquad [144]$$

If p_{31}, p_{32}, p_{33} are regarded as components of a vector p_3, and x_1, x_2, x_3 as those of a vector x, Eq. 144 demands that the vector x be orthogonal for the vector p_3, which, in the simpler case considered above, is coincident in direction with co-ordinate axis number 3. The constraint plane is normal to the vector p_3, and may be fixed at will through an appropriate choice of this vector.

The abridged quadratic form $\bar{F}$ is such a function of two variables that $\bar{F} = 1$ becomes the equation of the elliptic intersection of the original ellipsoid with the chosen constraint plane. Since the variables involved in $\bar{F}$ must refer to a two-dimensional orthogonal co-ordinate system lying within the constraint plane, it is clear that these variables are not simply

two of the original ones. To find $\bar{F}$ one must first determine a new set of orthogonal co-ordinate axes $1'$, $2'$, $3'$, such that one of these, say axis $3'$, is coincident with the vector p_3 (normal to the constraint plane). The new axes $1'$ and $2'$ will then lie in the constraint plane; and in terms of the new variables $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, referring to the new axes, it becomes clear that the constraint 144 is expressed by the simple relation $\bar{x}_3 = 0$. The problem is thus reduced to the simple form considered above. The abridged form $\bar{F}$ is found by first subjecting the original variables in F to the orthogonal transformation appropriate to changing from axes 1, 2, 3, to axes $1'$, $2'$, $3'$, and then dropping all terms involving $\bar{x}_3$.

The crux of the procedure lies in finding the proper axes $1'$, $2'$, $3'$, and from these the appropriate orthogonal transformation. Since axis $3'$ is coincident with the vector p_3 defined by the constraint Eq. 144, the problem is essentially that of associating with this vector p_3, two other vectors p_1 and p_2 such that the three together form a mutually orthogonal set. To carry out this procedure, one would begin by finding first a vector p_2 normal to p_3, and then a vector p_1 normal to the other two. Since there exists an infinite number of vectors p_2 normal to p_3, it is clear that the procedure as a whole is not unique, and there exists an infinite number of functions $\bar{F}$ of which any one may appropriately be called the abridged form of F for the stated constraint.

The details of the procedure just described are best made clear through a numerical example. Suppose one has given

$$F = x_1^2 + 2x_2^2 + 3x_3^2 \tag{145}$$

and the linear constraint

$$x_1 + x_2 + x_3 = 0 \tag{146}$$

Choosing F in the form of a sum of squares does not detract from the generality of the procedure to be discussed here.

The constraint vector p_3 has the components [1, 1, 1]. A procedure for finding vectors p_1 and p_2 such that the three vectors form a mutually orthogonal set may be patterned after the methods discussed in Art. 7, Ch. III. In this way the following matrix is readily found:

$$\mathcal{P} = \begin{bmatrix} -1 & -1 & 2 \\ 1 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{147}$$

the components of the vectors p_1, p_2, p_3 being defined, respectively, by the elements of the first, second, and third rows. As already mentioned, the determination of this matrix $\mathcal{P}$ is not unique. It may readily be checked by inspection that the three rows of $\mathcal{P}$ define a mutually orthogonal set of vectors.

Through dividing the components of each vector by the square root of the sum of the squares of its components (called normalization) one obtains a corresponding set of unit vectors, whose matrix is orthogonal (defines mutually orthogonal unit vectors by columns as well as by rows). This orthogonal matrix, which reads

$$\mathcal{O} = \begin{bmatrix} -0.408 & -0.408 & 0.817 \\ 0.707 & -0.707 & 0.0 \\ 0.577 & 0.577 & 0.577 \end{bmatrix} \qquad [148]$$

yields the desired transformation from the co-ordinates x_1, x_2, x_3 to $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, thus

$$\mathcal{O} \times x] = \bar{x}] \qquad [149]$$

Hence, according to discussion earlier in this chapter, the quadratic form in terms of the new variables has the matrix

$$\bar{\mathcal{Q}} = \mathcal{O} \times \mathcal{Q} \times \mathcal{O}_t \qquad [150]$$

where $\mathcal{Q}$ is the matrix of the given form.

In the numerical example considered here, one has

$$\bar{\mathcal{Q}} = \mathcal{O} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \times \mathcal{O}_t \qquad [151]$$

and with Eq. 148 this yields

$$\bar{\mathcal{Q}} = \begin{bmatrix} 2.5 & 0.2885 & 0.707 \\ 0.2885 & 1.5 & -0.408 \\ 0.707 & -0.408 & 2.0 \end{bmatrix} \qquad [152]$$

Deleting the third row and column, one has the matrix of the desired abridged quadratic form, which reads

$$\bar{F} = 2.5\bar{x}_1^2 + 0.577\bar{x}_1\bar{x}_2 + 1.5\bar{x}_2^2 \qquad [153]$$

In a more general case in which the given quadratic form F involves any number of dimensions (say n) and $t < n$ linear independent constraints are specified by the equations

$$\begin{aligned} p_{r1}x_1 + p_{r2}x_2 + \cdots + p_{rn}x_n &= 0 \\ p_{r+1,1}x_1 + p_{r+1,2}x_2 + \cdots + p_{r+1,n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_{n1}x_1 + p_{n2}x_2 + \cdots + p_{nn}x_n &= 0 \end{aligned} \qquad [154]$$

in which $r = n - t + 1$, one has given t vectors $p_r, p_{r+1}, \cdots p_n$ with components defined respectively by the coefficients in these equations taken by rows. The t planes orthogonal to these vectors are the constraint

planes, and the desired abridged quadratic form should have its variables confined to lie in the intersection common to these planes.

Geometrically it is now convenient to think of the t independent vectors $p_r, p_{r+1}, \cdots p_n$ as occupying a t-dimensional subspace immersed in the n-dimensional one. The common intersection of the constraint planes then occupies a $(n - t)$-dimensional subspace, and these two subspaces are mutually orthogonal to one another. One must find a new orthogonal set of co-ordinate axes such that the first $(n - t)$ of these lie in the $(n - t)$-dimensional subspace, while the remaining t of them lie in the t-dimensional subspace occupied by the vectors $p_r, p_{r+1}, \cdots p_n$. In this new co-ordinate system the constraint equations 154 become simply $\bar{x}_r = \bar{x}_{r+1} = \cdots = \bar{x}_n = 0$, and the variables $\bar{x}_1, \bar{x}_2, \cdots \bar{x}_{n-t}$ must lie in the common intersection of the constraint planes because these lie in the $(n - t)$-dimensional subspace.

The first step in the process of finding an appropriate set of new co-ordinate axes is to determine a mutually orthogonal set of t vectors that occupy the same subspace as the constraint vectors $p_r, p_{r+1}, \cdots p_n$. If the latter happen to be mutually orthogonal, this first step is already done (as may be said of the single constraint case discussed above), but in general the constraint vectors are any independent set, not necessarily an orthogonal one.

A mutually orthogonal set of vectors occupying the t-dimensional subspace is any set of t mutually orthogonal vectors each of which is expressible as a linear combination of the given constraint vectors. Such a set may be formed in an infinite variety of ways. One may begin, for example, by choosing p_r as the first of the desired set. Then one may determine a second vector, orthogonal to p_r, and expressed as a linear combination of p_r and p_{r+1}. Next one determines a third vector through demanding that it be orthogonal to the first two already found, and expressed as a linear combination of p_r, p_{r+1}, p_{r+2}, and so forth.

Specifically this process may be indicated as follows, letting the desired mutually orthogonal vector set be denoted by $\bar{p}_r, \bar{p}_{r+1}, \cdots \bar{p}_n$. To begin with, one chooses $\bar{p}_r = p_r$. Next

$$\begin{aligned} \bar{p}_{r+1} &= \alpha p_r + \beta p_{r+1} \\ \bar{p}_r \cdot \bar{p}_{r+1} &= \alpha p_r \cdot p_r + \beta p_r \cdot p_{r+1} = 0 \end{aligned} \qquad [155]$$

in which α and β are nonzero. Here β may be chosen arbitrarily. Thus letting $\beta = -1$ gives

$$\alpha = \frac{p_r \cdot p_{r+1}}{p_r \cdot p_r} \qquad [156]$$

and thus $\bar{p}_{r+1}$ is found. Then one writes

$$\begin{aligned} \bar{p}_{r+2} &= \alpha' p_r + \beta' p_{r+1} + \gamma' p_{r+2} \\ \bar{p}_r \cdot \bar{p}_{r+2} &= 0 \\ \bar{p}_{r+1} \cdot \bar{p}_{r+2} &= 0 \end{aligned} \qquad [157]$$

in which γ' may be chosen at will and the last two equations solved for α' and β', thus determining $\bar{p}_{r+2}$.

In the next step one will have to solve three equations simultaneously for three unknown coefficients, etc. If the number of constraints is large, the computations may become tedious, but they remain straightforward. There are, of course, other ways in which the desired set of t mutually orthogonal vectors $\bar{p}_r \cdots \bar{p}_n$ may be found, the method given here being rather simple in principle and no more tedious computationally than any other.

One must now associate with these t vectors, $n - t$ additional ones so as to obtain a complete set of n mutually orthogonal vectors $\bar{p}_1 \cdots \bar{p}_n$. This second step is carried out as in the single constraint case. The directions of these n vectors are those of the desired co-ordinate axes. The orthogonal transformation from the original co-ordinates to these new co-ordinate axes has a matrix $\mathfrak{O}$ whose vector set is $\bar{p}_1/|\bar{p}_1|$, $\bar{p}_2/|\bar{p}_2|$, $\cdots \bar{p}_n/|\bar{p}_n|$; that is, a set of unit vectors coincident with $\bar{p}_1 \cdots \bar{p}_n$. Equation 149 expresses the co-ordinate transformation, and Eq. 150 gives the matrix of the quadratic form F in the new variables $\bar{x}_k$ in terms of the matrix of F for the original variables x_k. In terms of the new variables

$$F = \underline{\bar{x}} \times \bar{\mathfrak{a}} \times \bar{x}] \qquad [158]$$

and the arbitrary linear constraints, expressed in terms of the original variables by Eqs. 154, are simply $\bar{x}_r = \bar{x}_{r+1} = \cdots = \bar{x}_n = 0$. The desired abridged quadratic form is that part of F remaining after terms involving the last t variables are discarded.

Perhaps it might be well to restate what has been done above in another way. Specification of the linear constraints 154 is equivalent to demanding that the vector x with components $x_1 \cdots x_n$ is no longer free to assume any orientation in space, but is required always to be *simultaneously* orthogonal to t arbitrary and independent vectors $p_r \cdots p_n$. Now these t vectors occupy only t of the n available dimensions of the given space. Hence, so long as the vector x moves about so as to stay outside the t-dimensional subspace occupied by the constraint vectors, it will fulfill the stated orthogonality restriction. For example, for $n = 3$ and $t = 1$, the vector x is free to move in a plane normal to the single specified con-

straint vector; for $n = 3$ and $t = 2$, the vector x is restricted to the one remaining dimension, defined as that direction in space which is orthogonal to the plane (t-dimensional subspace) determined by (or occupied by) the two constraint vectors.

When n is larger than 3, one must be mentally sufficiently adaptive in one's thought to comprehend the analogous geometry implied by continuing the identical algebraic reasoning beyond $n = 3$. For example, with $n = 4$ and $t = 2$, one must visualize a two-dimensional plane determined by the two constraint vectors as defining a corresponding two-dimensional subspace, and recognize that there are two other dimensions left over which define another subspace (the $n - t$ dimensional one in this case) orthogonal to the plane determined by the constraint vectors in the same sense that if the vector x remains in this second two-dimensional subspace, it remains simultaneously orthogonal to the stated constraint vectors.

Stated in terms of constraint planes, which are normal to their respective constraint vectors, one must recognize that the $(n - t)$-dimensional subspace to which the vector x is restricted should be interpreted as a resultant common intersection of these planes. Again for $n = 3$ and $t = 2$, the two constraint planes intersect in a line, so that only one dimension is left for the vector x to exist in. However, for $n = 4$ and $t = 2$, the "intersection" of the two constraint planes becomes a two-dimensional subspace. In general, one must visualize the possibility of t planes in an n-dimensional space as possessing a common intersection which is $(n - t)$-dimensional, the latter defining a subspace orthogonal to the one occupied by the constraint vectors.

What one is asked to do in the abridgment process is to find n mutually orthogonal unit vectors in the n-dimensional space such that t of these are linear combinations of the constraint vectors. These t vectors then clearly occupy the same subspace as the constraint vectors. Since the remaining $(n - t)$ vectors are simultaneously orthogonal to the first t, they must be simultaneously orthogonal to the constraint vectors, and hence they define the $(n - t)$-dimensional subspace in which the vector x can move and still conform with the restriction that it be orthogonal to all the constraint vectors.

If one chooses new co-ordinate axes coincident in direction with this set of mutually orthogonal unit vectors, the matrix of the linear transformation, expressing the co-ordinates with respect to the new axes in terms of those relating to the original axes according to Eq. 149, is clearly that orthogonal matrix having these unit vectors as its vector set. In terms of the new co-ordinates the original constraints are equivalent to setting equal to zero those variables that refer to the axes lying within the subspace occupied by the constraint vectors, because these simplified

constraint equations clearly demand no more nor no less than the original ones.

While there exists an infinite number of possible sets of new co-ordinate axes fulfilling the conditions just stated, note that the two subspaces defined by such axes are unique. For example, with $n = 3$ and $t = 2$, any co-ordinate system of which two axes lie in the plane determined by the constraint vectors (the t-dimensional subspace) is acceptable. This plane, however, is fixed, and so is the remaining direction normal to it (the $n - t$ dimensional subspace) no matter which one of the infinite possible choices one makes in determining a specific set of new axes.

In summary one may say that the discussions of this article show how one can convert an arbitrary set of t linear constraints, such as those expressed by the Eqs. 154, into an equivalent set having the simple form of demanding that t of the variables be zero. The latter will, of course, not be the original variables but new ones related to the original ones by an orthogonal transformation. In the next article the object is to investigate the effect of applied constraints upon the latent roots of a quadratic form. Since these latent roots are unchanged if the variables in the quadratic form are subjected to an orthogonal transformation, the form F has the same latent roots when it is expressed in terms of the new variables $\bar{x}_1 \cdots \bar{x}_n$ as it does when it is expressed in terms of the original variables $x_1 \cdots x_n$. One may therefore say that the effect upon the latent roots of F caused by imposing an arbitrary set of t linear constraints may be studied without loss in generality by considering only the simple case in which the constraints have the form expressed by setting t of the variables equal to zero.

15. The effect of constraints upon the latent roots of a quadratic form

In this discussion only the absolute values of the latent roots are of interest. When the quadratic form is positive definite, these roots are all positive real numbers, and there is no need to emphasize specifically the fact that only their absolute values are to be considered. It is only in the more general case in which the latent roots may have negative as well as positive real values that such a distinction is necessary. However, since the discussion in Art. 10 regarding the iterated forms shows that the iterated form of even order can have only positive latent roots (which are those of the original form raised to an even power), it is clear that, although the present discussion be restricted to the consideration of positive definite forms, the conclusions reached apply equally well to the magnitudes of the latent roots in the general case.

In Art. 5 it is shown that the latent roots $\lambda_1, \lambda_2, \cdots \lambda_n$ of the form

$$F = \sum_{i,k=1}^{n} a_{ik}x_i x_k \qquad [159]$$

may be regarded as the extrema of F subject to the restriction of the x_k's expressed by

$$\sum_{k=1}^{n} x_k^2 = 1 \qquad [160]$$

The correctness of this result may be made evident through considering the form F reduced to the normal form

$$F = \lambda_1\bar{x}_1^2 + \lambda_2\bar{x}_2^2 + \cdots + \lambda_n\bar{x}_n^2 \qquad [161]$$

by means of the orthogonal transformation

$$x] = \mathfrak{L}\bar{x}] \qquad [162]$$

involving the modal matrix $\mathfrak{L}$. Because of the orthogonality of the latter transformation, the condition 160 is unchanged in form, that is,

$$\sum_{k=1}^{n} \bar{x}_k^2 = 1 \qquad [163]$$

In terms of F as given by Eq. 161, and the condition expressed by Eq. 163, it is clear by inspection that the latent roots are extrema of F. More specifically, if the roots are numbered in such a way that

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_n \qquad [164]$$

one may see that the largest value of F for the condition 163 occurs for $\bar{x}_1 = 1$, $\bar{x}_2 = \bar{x}_3 = \cdots = \bar{x}_n = 0$, and equals λ_1. The next largest stationary value of F under the same condition occurs for $\bar{x}_2 = 1, \bar{x}_1 = \bar{x}_3 = \cdots = \bar{x}_n = 0$, and equals λ_2; and so forth.

If the variables $x_1 \cdots x_n$ in F are subjected to an arbitrary set of t independent linear constraints, the resulting abridged form $\bar{F}$ in $n - t$ variables has $n - t$ latent roots which may be numbered so as to conform to the sequence

$$\bar{\lambda}_1 > \bar{\lambda}_2 > \bar{\lambda}_3 > \cdots > \bar{\lambda}_{n-t} \qquad [165]$$

It is the object of the following discussion to establish relations between the magnitudes of the latent roots of F and those of $\bar{F}$.

The constraint equations, which are assumed to have the form

$$x_1 = x_2 = x_3 = \cdots = x_t = 0 \qquad [166]$$

are alternately expressed in terms of the variables $\bar{x}_k$ (related to the x_k

through the orthogonal transformation 162) as

$$\sum_{k=1}^{n} l_{sk}\bar{x}_k = 0 \qquad (s = 1, 2, \cdots t) \tag{167}$$

in which the l_{sk} are elements of $\mathfrak{L}$.

Although F, subject to the linear constraints, has latent roots that are different from $\lambda_1 \cdots \lambda_n$, nevertheless the expression 161 for F may be used to compute values of F corresponding to *any* values of the variables $\bar{x}_1 \cdots \bar{x}_n$, and hence yields values for the abridged quadratic form provided only that the assumed values for $\bar{x}_1 \cdots \bar{x}_n$ conform to the constraint equations 167. If in these equations one chooses to let $\bar{x}_{t+2} = \bar{x}_{t+3} = \cdots = \bar{x}_n = 0$, there results a set of t equations in $t + 1$ unknowns, which surely possess a nontrivial solution for $\bar{x}_1 \cdots \bar{x}_{t+1}$ in agreement with the condition 163. For such a set of $\bar{x}_k$-values, Eq. 161 yields

$$F = \lambda_1\bar{x}_1^2 + \lambda_2\bar{x}_2^2 + \cdots + \lambda_{t+1}\bar{x}_{t+1}^2 \geqq \lambda_{t+1}(\bar{x}_1^2 + \cdots \bar{x}_{t+1}^2) = \lambda_{t+1} \tag{168}$$

as may be appreciated by noting the inequalities in 164 and the condition 163.

Since F, subject to the linear constraints 166 or 167, is the abridged form $\bar{F}$, the result 168 shows that a possible value of the abridged form is at least as large as the latent root λ_{t+1} of the corresponding unabridged form. If the maximum-minimum problem discussed in Art. 5 is applied to the abridged form $\bar{F}$, one observes that the maximum value of $\bar{F}$, subject to a condition in terms of its variables similar to 160 or 163, is $\bar{\lambda}_1$. The next largest stationary value is $\bar{\lambda}_2$, and so forth. Inasmuch as it has been shown that a possible value for $\bar{F}$ is at least as large as λ_{t+1}, it is clear that the largest value $\bar{\lambda}_1$ is surely as large as this, that is,

$$\bar{\lambda}_1 \geqq \lambda_{t+1} \tag{169}$$

Suppose now that v additional constraints are imposed upon F, making $t + v$ constraints in all. The resulting abridged form, which may be denoted by $\bar{\bar{F}}$, can also be regarded as resulting from imposing the set of v constraints upon $\bar{F}$. If the latent roots of $\bar{\bar{F}}$ are denoted and numbered according to the sequence

$$\bar{\bar{\lambda}}_1 > \bar{\bar{\lambda}}_2 > \bar{\bar{\lambda}}_3 > \cdots > \bar{\bar{\lambda}}_{n-t-v} \tag{170}$$

then one may write two additional relations similar to 169 which read

$$\bar{\bar{\lambda}}_1 \geqq \lambda_{t+v+1} \tag{171}$$

and

$$\bar{\bar{\lambda}}_1 \geqq \bar{\lambda}_{v+1} \tag{172}$$

In any one of the relations 169, 171, or 172 the equals sign holds only

if the constraints are chosen in a particular way. Thus for a particular form of the v constraints, the equals sign in 172 may be assumed to hold, but it will not simultaneously hold in 171 also as long as the original t constraints are considered to be arbitrary. Thus the relations 171 and 172 are seen to yield

$$\bar{\lambda}_{v+1} \geqq \lambda_{t+v+1} \tag{173}$$

which establishes relations between all the latent roots of $\bar{F}$ and a like number of those of F, since the integer v can have any value from 0 to $n - t - 1$.

Considering again the interpretation of the latent roots $\bar{\lambda}_1 \cdots \bar{\lambda}_{n-t}$ of $\bar{F}$ in the manner that those of F are interpreted in Art. 5, bearing in mind that $\bar{F}$ is the result of imposing certain constraints upon F, and hence that all the extrema of $\bar{F}$ are smaller than the respective ones of F, which are attainable only if its variables are free to assume particular sets of values, one recognizes the following additional relation as being true:

$$\lambda_s \geqq \bar{\lambda}_s \qquad (s = 1, 2, \cdots n - t) \tag{174}$$

Together with 173, one may summarize the results so far in the form

$$\lambda_s \geqq \bar{\lambda}_s \geqq \lambda_{s+t} \tag{175}$$

in which the index s may be given the integer values $1, 2, \cdots n - t$. A useful relation between the $n - t$ latent roots of $\bar{F}$ in terms of the n latent roots of F is thus established.

A case of particular interest is that in which a single constraint is imposed upon F. The latent roots of F and $\bar{F}$ are then related as expressed by

$$\lambda_1 \geqq \bar{\lambda}_1 \geqq \lambda_2 \geqq \bar{\lambda}_2 \geqq \lambda_3 \geqq \cdots \geqq \bar{\lambda}_{n-1} \geqq \lambda_n \tag{176}$$

These inequalities are sometimes referred to as expressing the *separation property* of the roots of F and $\bar{F}$.

An interesting geometrical view of this last result is had through visualizing the ellipsoid associated with F, and the constraint plane as slicing the ellipsoid centrally at an arbitrary angle. The intersection of the ellipsoid with the plane yields an ellipse with principal axes whose lengths are clearly intermediate as compared with those of the ellipsoid.

With reference to the pair of forms F_1 and F_2, discussed in Arts. 11 and 12, it may be pointed out that the results of the present article apply also to the roots $\lambda^{(a/b)}$ and those of an abridged pair of forms resulting from subjecting their common variables $x_1 \cdots x_n$ to an arbitrary set of independent linear constraints.

PROBLEMS

1. Write down the matrices of the following quadratic forms and determine whether or not they are singular.

$$F_1 = 5x_1^2 - 8x_1x_2 + 3x_1x_3 + 4x_1x_4 + 2x_2^2 + 4x_2x_3 - 2x_2x_4 + 5x_3^2 + 2x_3x_4 + 6x_4^2$$

$$F_2 = 8x_1^2 + 3x_1x_2 + 7x_1x_3 + 9x_2x_1 + 9x_2^2 - 3x_2x_3 - 3x_3x_1 + 5x_3x_2 + x_3^2$$

2. Show that the matrix of a quadratic form can be written as

$$\mathfrak{A} = \frac{1}{2}\left[\frac{\partial^2 F}{\partial x_i\,\partial x_k}\right] \qquad i, k = 1, 2, \cdots n$$

3. Transform congruently the matrices

$$\begin{bmatrix} 4 & 2 & 1 & 0 \\ 2 & 6 & 4 & -1 \\ 1 & 4 & 8 & 3 \\ 0 & -1 & 3 & 10 \end{bmatrix} \quad \begin{bmatrix} 5 & 3 & -2 \\ 3 & 10 & -1 \\ -2 & -1 & 15 \end{bmatrix} \quad \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix}$$

respectively with

$$\begin{bmatrix} 5 & 3 & 4 & 1 \\ 3 & 8 & 1 & 1 \\ 0 & 3 & 5 & 0 \\ 1 & 2 & -1 & 4 \end{bmatrix} \quad \begin{bmatrix} 2 & 2 & -1 \\ 3 & 5 & 7 \\ 2 & 3 & 1 \end{bmatrix} \quad \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$$

4. For what type of transformation matrix is the collineatory transformation identical with a congruent one? Out of the following matrices, pick by inspection those that have this property:

$$\begin{bmatrix} 2 & 1 & -1 \\ 3 & 5 & -2 \\ 1 & 0 & 3 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{3} & \frac{4}{3\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{3} & -\frac{2}{3\sqrt{5}} \\ 0 & \frac{2}{3} & \frac{\sqrt{5}}{3} \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & -2 \\ 1 & 4 & 1 \\ 1 & -4 & 1 \end{bmatrix}$$

Check your selection by carrying out the transformation upon the matrix

$$\begin{bmatrix} 5 & 3 & -2 \\ 3 & 10 & -1 \\ -2 & -1 & 15 \end{bmatrix}$$

5. Compute the latent roots and proper values of the following quadratic forms:

$$F_1 = \tfrac{1}{45}[83x_1^2 + 52x_1x_2 - 40x_1x_3 + 122x_2^2 + 20x_2x_3 + 65x_3^2]$$

$$F_2 = \tfrac{1}{45}[21x_1^2 + 24x_1x_2 + 120x_1x_3 + 39x_2^2 - 60x_2x_3 + 30x_3^2]$$

$$F_3 = \tfrac{1}{45}[49x_1^2 - 4x_1x_2 - 80x_1x_3 + 46x_2^2 + 40x_2x_3 + 40x_3^2]$$

$$F_4 = x_1^2 + 2\sqrt{3}x_1x_2 - x_2^2$$

6. Find the lengths of the principal semiaxes of the ellipsoid represented by the quadratic form:

$$F = 177x_1^2 + 228x_1x_2 - 120x_1x_3 + 348x_2^2 + 60x_2x_3 + 105x_3^2 = 45$$

7. Compute the modal matrix of the quadratic form given in Prob. 6 and make an isometric plot showing the positions of the principal axes.

8. Show that the quadratic surface

$$2x_1^2 + 5x_2^2 - 6x_2x_3 + 5x_3^2 = 2$$

represents an ellipsoid of revolution. Compute the lengths of the semiaxes and find their direction cosines with reference to the given co-ordinate system. Write down the modal matrix. Is it unique?

9. Reduce the following quadratic forms to sums of squares through orthogonal transformation and give the respective transformation matrices:

$$F_1 = 3.5x_1^2 - 2x_1x_2 + x_1x_3 + 1.6x_2^2 + 0.4x_2x_3 + 1.9x_3^2$$

$$F_2 = \tfrac{1}{153}[130x_1^2 - 284x_1x_2 - 188x_1x_3 + 181x_2^2 - 256x_2x_3 - 5x_3^2]$$

$$F_3 = -2.6x_1^2 + 2.4x_1x_2 + 0.6x_2$$

$$F_4 = 0.4(x_1^2 + \sqrt{3}x_1x_2 + 2x_2^2)$$

10. Reduce the following quadratic term to its normal form:

$$F = 0.487x_1^2 - 0.784x_1x_2 + 0.632x_1x_3 + 0.445x_1x_4 + 0.487x_2^2 + 0.632x_2x_3 + 0.445x_2x_4 + 1.892x_3^2 + 0.111x_4^2$$

What is its rank?

11. Find the orthogonal matrix which transforms

$$F = 0.392x_1^2 - 0.948x_1x_2 - 0.632x_1x_3 + 0.444x_1x_4 + 0.392x_2^2 - 0.632x_2x_3 - 0.444x_2x_4 + 1.892x_3^2 + 0.111x_4^2$$

to the normal form. Find the latent roots and the signature.

12. By forming the Gramian determinant, establish the linear dependence or independence of the vector set defined by the row matrices:

$$[1\ \ 3\ \ 4\ \ 5\ \ 1\ \ 3] \qquad [2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7] \qquad [1\ \ 6\ \ 8\ \ 10\ \ {-3}\ \ 2]$$

Repeat for the set

$$[1\ \ {-1}\ \ 4\ \ 8\ \ 2] \qquad [3\ \ {-3}\ \ {-2}\ \ {-1}\ \ 0] \qquad [4\ \ 2\ \ 5\ \ 8\ \ {-8}]$$

13. By means of elementary transformations find a matrix $\mathcal{Q}$ which will reduce to a sum of squares the quadratic form

$$F = 4x_1^2 + 5x_1x_2 - \tfrac{5}{2}x_1x_3 + 7x_2^2 + \tfrac{3}{2}x_2x_3 + 2x_3^2$$

through a congruent transformation of its matrix $\mathcal{A}$. Determine the corresponding diagonal matrix $\mathcal{D}_1$ and the signature.

Through a similar procedure find another matrix $\mathcal{P}$ which also yields a reduction to a sum of squares. Express the relation between the two diagonal forms $\mathcal{D}$ and $\mathcal{D}'$ of the reduced matrices by computing the diagonal matrix Δ in the equation $\mathcal{D}' = \Delta_t\mathcal{D}\Delta$. Show that it is always possible to find such a real Δ relating the alternate diagonal forms of a symmetric matrix.

14. Using the following vector set,

$$[1\ \ 1\ \ {-1}] \qquad [-3\ \ 3\ \ 5] \qquad [-5\ \ 0\ \ 5]$$

form a matrix $\mathfrak{A}$ whose elements are computed as are those in the Gramian determinant. Find the latent roots of $\mathfrak{A}$ and check the positive character of the principal minors of its determinant.

15. If possible, reduce each of the following quadratic forms to its canonic form and give the corresponding modal matrix as well as the transformation required to pass from the normal to the canonic forms.

$$F_1 = [83x_1^2 + 52x_1x_2 - 40x_1x_3 + 122x_2^2 + 20x_2x_3 + 65x_3^2]45^{-1}$$

$$F_2 = [49x_1^2 - 4x_1x_2 - 80x_1x_3 + 46x_2^2 + 40x_2x_3 + 40x_3^2]45^{-1}$$

16. Determine whether or not the following quadratic form is positive definite:

$$F = 3x_1^2 - 4x_1x_2 - 4x_1x_3 + 5x_2^2 + 2x_2x_3 + 4x_3^2$$

17. Find the triangular matrix $\mathcal{P}$ that generates the matrix $\mathfrak{A}$ of the quadratic form

$$F = x_1^2 - 6x_1x_2 - 2x_1x_3 + 4x_1x_4 + 13x_2^2 + 10x_2x_3 - 20x_2x_4 + 11x_3^2 - 14x_3x_4 + 25x_4^2$$

and check the relation $\mathfrak{A} = \mathcal{P}_t\mathcal{P}$.

18. (a) Find the extrema of the quadratic form

$$F = \tfrac{1}{4}[-7x_1^2 - 10\sqrt{3}x_1x_2 + 3x_2^2]$$

when the variables x_1, x_2 are subjected to the condition that the point $P(x_1,x_2)$ shall lie on the unit circle.

(b) Find the extrema of

$$F = \tfrac{1}{8}[4x_1^2 + 4x_1x_2 + 4x_1x_3 + 149x_2^2 - 278x_2x_3 + 149x_3^2]$$

subject to the condition that $P(x_1,x_2,x_3)$ shall lie on a sphere of radius 2. Determine the signature of this quadratic form.

19. Let $\mathfrak{A}$ be the matrix of a positive definite quadratic form, A its discriminant, and $\mathcal{P}$ a triangular matrix which generates $\mathfrak{A}$ according to the relation $\mathfrak{A} = \mathcal{P}_t\mathcal{P}$. Show that a possible procedure for the formation of $\mathcal{P}$ is given by the relations: $p_{kj} = 0$ for $k > j$; $p_{kj} \neq 0$ for $k \leq j$.

$$p_{1j} = \pm\frac{a_{1j}}{\sqrt{a_{11}}} \qquad p_{2j} = \pm\frac{B_{1j}}{\sqrt{a_{11}B_{12}}} \qquad p_{3j} = \pm\frac{C_{12j}}{\sqrt{B_{12}C_{123}}} \cdots$$

$$p_{kj} = \pm\frac{K_{1\cdots(k-1)j}}{\sqrt{J_{1\cdots(k-1)}K_{1\cdots k}}} \qquad p_{nn} = \pm\frac{\sqrt{A}}{\sqrt{M_{1\cdots(n-1)}}}$$

Here a_{1j} are the elements of the first row of $\mathfrak{A}$. The B_{1j} are second-order minors formed from the first two rows of $\mathfrak{A}$ by selecting the first and the jth columns. The C_{12j} are formed from the first three rows of $\mathfrak{A}$ by selecting columns 1, 2, and j; and so forth.

20. By using the procedure given in the preceding problem compute a triangular matrix $\mathcal{P}$ which generates the matrix of the quadratic form given in Prob. 17.

21. Show through a proper expansion of the characteristic determinant of a matrix that its characteristic equation can be written in the form

$$
\begin{aligned}
(-\lambda)^n + (-\lambda)^{n-1} &\textstyle\sum \text{(principal diagonal elements)} \\
+ (-\lambda)^{n-2} &\textstyle\sum \text{(principal minors of order 2)} \\
+ (-\lambda)^{n-3} &\textstyle\sum \text{(principal minors of order 3)} \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
-\lambda &\textstyle\sum \text{(principal minors of order } n-1) \\
&+ \text{determinant} = 0
\end{aligned}
$$

In view of this result, show that the sum of the latent roots of a matrix equals the sum of its principal diagonal elements, and that the product of the latent roots equals the determinant. Show further that if the determinant is zero, at least one latent root is zero; if the matrix is of rank $n - 2$, at least two latent roots are zero; and so on.

22. Using the result of the previous problem, show that the roots of the equation $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0$ are the latent roots of the matrix

$$
\begin{bmatrix}
0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 1 & \cdots & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & & \cdots & 0 & 1 \\
-a_0 & -a_1 & -a_2 & & \cdots & \cdots & -a_{n-1}
\end{bmatrix}
$$

23. Let $\mathcal{A}$ be an nth order square matrix and $\lambda_1, \lambda_2, \cdots \lambda_n$ its latent roots. The mth power of $\mathcal{A}$ is denoted by $\mathcal{B}$.

(a) If the latent roots are real and distinct, and numbered so that $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, show that for a sufficiently large m:

$$
\begin{aligned}
|\lambda_1|^m &\cong \textstyle\sum \text{(principal diagonal elements of } B) \\
|\lambda_1\lambda_2|^m &\cong \textstyle\sum \text{(principal minors of order 2)} \\
|\lambda_1\lambda_2\lambda_3|^m &\cong \textstyle\sum \text{(principal minors of order 3)} \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
|\lambda_1\lambda_2\cdots\lambda_n|^m &\cong B
\end{aligned}
$$

(b) If the latent roots are real but there are α repeated roots so that

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_g| > |\lambda_{g+1}| = |\lambda_{g+2}| = \cdots = |\lambda_{g+\alpha}| > |\lambda_{g+\alpha+1}| > \cdots > |\lambda_n|$$

derive the relations

$$|\lambda_j|^m \cong \frac{\sum \text{(principal minors of order } j)}{\sum \text{(principal minors of order } j-1)} \qquad j = 1, 2, \cdots g; g+\alpha+1, g+\alpha+2, \cdots n$$

$$|\lambda_{g+1}|^m \cong \frac{\sum \text{(principal minors of order } g+1)}{\alpha \sum \text{(principal minors of order } g)}$$

(c) Suppose, in the sequence $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, that the first $h - 1$ roots are real but that the roots λ_h and λ_{h+1} are conjugate complex with the angle ϕ_h. In this case show that

$$|\lambda_h|^{2m} \cong \frac{\sum \text{(principal minors of order } h+1)}{\sum \text{(principal minors of order } h-1)}$$

and

$$2\,|\lambda_h|^m \cos m\phi_h \cong \frac{\sum \text{(principal minors of order } h)}{\sum \text{(principal minors of order } h-1)}$$

24. If F is a positive definite quadratic form in n variables, show that its characteristic determinant, when expanded as a polynomial in λ, consists of $n + 1$ terms with alternate algebraic signs, and that the same polynomial with all terms alike in sign is associated with a negative definite form.

25. Consider the quadratic surface $\sum_{i,k=1}^{n} a_{ik}x_ix_k = 1$ and the plane

$$\sum_{h=1}^{n} A_hx_h = 0$$

(a) Show that the intersection of the plane with the surface can be expressed as a quadratic surface of $n - 1$ dimensions in the form

$$\sum_{r,s=1}^{n-1} b_{rs}x'_rx'_s = 1$$

and compute the values of the coefficients b_{rs}.

(b) As a numerical illustration find the equation of the ellipse given by the intersection of the ellipsoid of revolution

$$\frac{x_1^2}{a^2} + \frac{x_2^2 + x_3^2}{b^2} = 1 \text{ with } a = \sqrt{2}, \qquad b = 1$$

and the plane

$$-2x_1 + 3x_2 + x_3 = 0$$

Find the lengths of the semiaxes of this ellipse and compute their direction cosines.

26. Given the quadratic surface

$$\sum_{i,k=1}^{n} a_{ik}x_ix_k = 1$$

and the parametric equation of a straight line

$$x_j = \alpha_jt + \beta_j \qquad j = 1, 2, \cdots n$$

in which α_j and β_j are constants and t is a variable parameter.

(a) Demonstrate that the line cuts the surface at most in two points and that the corresponding parameter values must satisfy the equation

$$\left(\sum_{i,k=1}^{n} a_{ik}\alpha_i\alpha_k\right)t^2 + 2\left(\sum_{i,k=1}^{n} a_{ik}\alpha_i\beta_k\right)t = 1 - \sum_{i,k=1}^{n} a_{ik}\beta_i\beta_k$$

Discuss the conditions under which there are 2, 1, or no points of intersection.

(b) Compute the co-ordinates x'_k and x''_k $(k = 1, 2, \cdots n)$ of the points of intersection.

(c) If the straight line passes through the origin, show that

$$x'_k = -x''_k = \frac{\alpha_k}{\sqrt{\sum_{i,k=1}^{n} a_{ik}\alpha_i\alpha_k}}$$

and interpret this result according to whether the quadratic form is positive definite or not.

(d) Write down the corresponding expressions appropriate to a quadratic form representing a cone.

27. Carry through the simultaneous reduction to the normal form of both quadratic forms in the following pairs:

(i) $F_1 = 2x_1^2 + 2x_1x_2 + x_2^2 \qquad F_2 = x_1^2 + 2x_1x_2 - x_2^2$

(ii) $F_1 = 2x_1^2 + 2x_1x_2 + 2x_1x_3 + 3x_2^2 - 2x_2x_3 + 2x_3^2$

$F_2 = x_1^2 + 3x_2^2 - 2x_2x_3 + 2x_2^2$

(iii) $F_1 = 2x_1^2 + 6x_1x_2 + 5x_2^2 \qquad F_2 = 2x_1x_2$

28. Compute the matrices $\mathfrak{L}^{(b)}$, $\Lambda_b^{1/2}$, and $\mathcal{P}$ (see Eq. 101), and the matrices $\mathfrak{A}^{(b)}$ and $\mathfrak{L}^{(a/b)}$ for each of the three reductions in Prob. 27.

29. Let $a_{11}x_1^2 + a_{22}x_2^2 = 1$ be the equation of an ellipse when referred to a system of oblique axes making an angle ϕ.

Show that the principal axes of the ellipse make angles α_1 and α_2 with the x_1-axis that are determined by the relation

$$\tan\alpha_{1,2} = \frac{(m^2 - 1) \pm \sqrt{(m^2 - 1)^2 + 4m^2\cos^2\phi}}{(m^2 + 1) \pm \sqrt{(m^2 + 1)^2 - 4m^2\sin^2\phi}}$$

in which $m = a_{11}/a_{22}$; and that these axes can have any desired orientations through the proper choice of the ratio m. Observe that the angle between the axes is 90°.

Write down the expression for a family of ellipses for which one principal axis coincides with the x_2-axis.

30. Consider the pair of quadratic forms:

$$F_1 = 3x_1^2 - 2\sqrt{2}x_1x_2 + 2x_2^2 \qquad \text{and} \qquad F_2 = x_1^2 + 2\sqrt{2}x_1x_2 - 2x_2^2$$

Regard the variables x_1 and x_2 as the contravariant co-ordinates of a point P referred to a system of axes for which the fundamental metric tensor has the matrix

$$\mathcal{G} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

(a) Make a plot of the curves $F_1 = 1$ and $F_2 = 1$ in this system of reference co-ordinates.

(b) Find the transformation of variables which simultaneously reduces both quadratic forms to their normal forms.

(c) Find the orientations of the new co-ordinate axes relative to the old ones.

(d) Compute the transformed fundamental matrix $\mathcal{G}$.

(e) Find the orientations of the principal axes of the resultant ellipse with respect to the new co-ordinate system.

31. Consider the pair of quadratic forms

$$F_1 = x_1^2 - 2x_1x_2 + 2x_1x_3 + 2x_2^2 - 2x_2x_3 + 3x_3^2$$

$$F_2 = 3x_1^2 - 2x_1x_2 + 2x_1x_3 - x_2^2 - 2x_2x_3 + x_3^2$$

in which x_1, x_2, x_3 are contravariant variables with respect to an oblique co-ordinate system consisting of a triad of axes whose mutual angles are all 60°.

Reduce both F_1 and F_2 to their normal forms through a transformation to a new co-ordinate system and find the orientations of the new co-ordinate axes with respect to the old ones.

32. Find the equation of the intersecting curve between the quadratic surface

$$83x_1^2 + 52x_1x_2 - 40x_1x_3 + 122x_2^2 + 20x_2x_3 + 65x_3^2 = 45$$

and the plane

$$5x_1 + 4x_2 - 3x_3 = 0$$

and compute the lengths of the semiaxes of this curve as well as their orientation with respect to the assumed rectangular Cartesian co-ordinates.

33. For the ellipsoid given in Prob. 32 find equations of the three planes containing the principal axes taken in pairs.

34. Find the maximum perpendicular distance from points on the ellipsoid of Prob. 32 to the plane

$$-2x_1 + x_2 + 2x_3 = 0$$

35. Given

$$F = \tfrac{1}{15}[21x_1^2 + 24x_1x_2 + 120x_1x_3 + 39x_2^2 - 60x_2x_3 + 30x_3^2]$$

Find the abridged quadratic form that results if the variables are subjected to the condition that they determine a point which is constrained to lie on a plane normal to a vector with the components 1, 2, -1, and compute the latent roots of this abridged form.

36. The rectangular Cartesian co-ordinates forming the reference system for the ellipse

$$3x_1^2 + 2x_2^2 + x_3^2 = 1$$

are subjected to a transformation to a new set of axes defined by the orthogonal vector set

$$\begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{3} & \frac{4}{3\sqrt{5}} \end{bmatrix} \quad \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{3} & -\frac{-2}{3\sqrt{5}} \end{bmatrix} \quad \begin{bmatrix} 0 & \frac{2}{3} & \frac{\sqrt{5}}{3} \end{bmatrix}$$

Find the equation of this ellipse in terms of the new variables and make plots in the new co-ordinate planes of the intersections of the ellipse with these planes.

37. The real quadratic form

$$F = \sum_{i,k=1}^{n} a_{ik}x_ix_k$$

is subjected to the set of t linear constraints

$$\sum_{k=1}^{n} p_{rk}x_k = 0 \quad r = 1, 2, \cdots t$$

and the condition

$$\sum_{k=1}^{n} x_k^2 = 1$$

In terms of the function

$$F \equiv G = F + \lambda\left(1 - \sum_{k=1}^{n} x_k^2\right) + 2\sum_{r=1}^{t}\sum_{k=1}^{n} \mu_r p_{rk}x_k$$

in which λ and $\mu_1 \cdots \mu_t$ are Lagrangian multipliers, show that the conditions for an

extremum of F subject to the constraints lead to the equations

$$\sum_{i=1}^{n} a_{ik}x_i - \lambda x_k + \sum_{r=1}^{t} \mu_r p_{rk} = 0 \qquad (k = 1, 2, \cdots n)$$

$$\sum_{k=1}^{n} p_{rk}x_k = 0 \qquad (r = 1, 2, \cdots t)$$

$$\sum_{k=1}^{n} x_k^2 = 1$$

Show that this system of equations is sufficient to determine the unknowns $x_1 \cdots x_n$, $\mu_1 \cdots \mu_t$, λ, and that the extremum of F is given by $F = \lambda$ in which the λ-values are roots of the equation

$$\begin{vmatrix} (a_{11} - \lambda) & a_{12} & \cdots & a_{1n} & p_{11} & \cdots & p_{t1} \\ a_{21} & (a_{22} - \lambda) & \cdots & a_{2n} & p_{12} & \cdots & p_{t2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & (a_{nn} - \lambda) & p_{1n} & \cdots & p_{tn} \\ p_{11} & p_{12} & \cdots & p_{1n} & 0 & \cdots & 0 \\ p_{21} & p_{22} & \cdots & p_{2n} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{t1} & p_{t2} & \cdots & p_{tn} & 0 & \cdots & 0 \end{vmatrix} = 0$$

38. Show that the determinant given in Prob. 37 furnishes the same latent roots as those corresponding to the abridged quadratic form. Following a method similar to that given in Prob. 21, obtain expressions for the coefficients of the various powers of λ in this characteristic equation.

39. Illustrate the procedure outlined in Prob. 37 with the quadratic form and the single constraint given in Prob. 32 and compare the results with the solution to that problem.

CHAPTER V

Vector Analysis

1. Preliminary remarks and definitions

The quantities considered in this chapter are functions of the co-ordinates of ordinary (three-dimensional) space. In some cases, they also may be functions of other independent variables such as the time.

A *scalar* is a function which, for each set of values of its independent variables, is completely characterized by a corresponding magnitude. If the function is defined for all points within a given region, it is there said to constitute a scalar field. Potential functions, such as the scalar potential in an electric field or the thermodynamic potential of an ideal gas, are common examples. The geographical altitude as a function of latitude and longitude is a two-dimensional example of a scalar field.

A *vector* is a function which is characterized at each point in space by means of a magnitude and a direction. If the function is defined for all points within a given region, it is there said to constitute a *vector field.* The earth's gravitational field of force or the velocity field of a fluid are familiar examples. The magnitude of a vector function is a scalar. The vector function may, therefore, be thought of as a scalar to which a direction is assigned at each point in space.

More specifically, however, two kinds of vector functions are distinguished according to their processes of derivation. Thus, for example, the *gradient* of a scalar potential function is a vector. A simple example is the gradient in a mountainous terrain. A vector of a physically different nature is that used to represent a mechanical torque. The torque is produced by a force acting upon a lever arm, and the resulting vector (by convention) stands normal to the plane determined by the force and the arm; that is, it coincides with the axis of rotation. The direction of the torque vector, moreover, must be defined in accordance with a right- or a left-hand screw rule.

These two types of vectors, such as a gradient and a torque, are distinguished respectively by the adjectives *polar* and *axial.* This distinction is not merely a superficial one which may be disposed of by the simple process of propounding a pair of suitable adjectives. One reason for making such a distinction is brought to light when the variables of these vector functions are subjected to a co-ordinate transformation, such as changing from a right-hand to a left-hand system of rectangular axes (see Fig. 1). In this case, the algebraic sign of the axial vector function is reversed and that of the polar vector is not. If both types of vector func-

tions are involved in a given problem, this circumstance must be carefully considered.

An axial vector may be the result of a vector product formed from two given vectors. It must be observed, however, that this is the case *only* if *both* the given vectors are either axial or polar. The vector product formed from a polar and an axial vector is polar.

Since a scalar may be the result of a scalar product of two vectors, it appears that this question regarding the distinction between two types

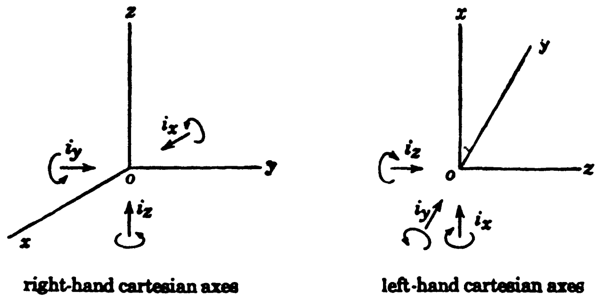


FIG. 1. Two systems of cartesian axes.

of vectors is not confined to vector functions. Thus a scalar function which results from the scalar product of a polar and an axial vector has different mathematical properties from those of a scalar function which is the scalar product of two polar or two axial vectors. The first of these functions reverses its algebraic sign when subjected to a transformation from a right- to a left-hand co-ordinate system; the second does not. The latter is invariant to any co-ordinate transformation, as a true scalar should be. An energy function is a scalar of this type. The other kind of scalar function, which is also encountered in physical problems, is called a *pseudoscalar*, since it has all the properties of a scalar except that it is not invariant to certain types of co-ordinate transformations.

The product of a scalar and a vector yields a vector of the same type. Multiplication with a pseudoscalar, however, changes an axial vector into a polar one, and vice versa. It should also be observed that the addition or subtraction of vectors or scalars should not be carried out without regard to their type or origin.

In order that the geometrical visualization of it may be facilitated, a field is commonly pictured as associated with a system of so-called *flow lines*. In hydrodynamics, for example, these are the paths which are traversed by the component particles of the fluid. Since the velocity vector is tangent to these trajectories at every point within the flow

region, certain physical characteristics of the vector field itself can be recognized from such a system of flow lines (*flow map*).*

The magnitude of the function representing the vector field at any point is given by the density of flow lines in the surface normal to the vector at that point. The number of lines chosen to represent unit density is, of course, arbitrary, but it is significant to observe that since the magnitude is a continuous function (except at certain points or surfaces where lines begin or end), *all* points within the field must be thought of as occupied by lines.

A given bunch of these lines, moreover, cannot intertwine with each other, because their continuous distribution would then necessitate a crossing of lines at some points, and this is impossible since their directions must everywhere be unique. The lines defining the longitudinal surface of a given bunch form what is known as a *tube*. Unless such a tube contains regions from which lines emanate or upon which they terminate, the total number of lines enclosed by it must evidently remain constant throughout the flow region defined by the tube.

From this more physical point of view, vector fields are distinguished according to either of two characteristically different (yet in a sense complementary) properties which they may separately or simultaneously possess. Thus if the flow map exhibits lines which close upon themselves (are endless), the field is said to exhibit *turbulent* characteristics. If none of the flow lines close upon themselves, the field is said to be *nonturbulent*.† In connection with the latter statement it must be recognized, of course, that a flow map for a finite region may exhibit no closed paths, yet the greater field may be turbulent, for some of the paths may close outside the finite mapped region.

A vector field which is solely turbulent (alternatively called *rotational* or *solenoidal*‡) is associated with a flow map containing closed paths only. Figure 2 shows an example of such a field. A nonturbulent field (also called an *irrotational* or a *potential* field) is associated with a flow map in which all the lines begin at a *source* and end upon a *sink* (or negative source). For this reason, the potential field is sometimes referred to as a source field, and the turbulent one as a source-free field. Figure 3 illustrates an irrotational field.

In the sense that a source is the cause or origin of any field, the turbulent field must, of course, also have its sources. These, however, are

*These matters are discussed in greater detail in *Electric Circuits*, pp. 23–71.

†The term *lamellar* is sometimes used to describe such a field.

‡A solenoid is a channel or tube. The term " solenoidal " does not appear to be particularly appropriate because tubes of flow lines can also be mapped in a potential field. The important characteristic of the rotational field is the fact that its tubes close upon themselves, each forming an endless conduit containing the same number of flow lines throughout the length of its circuit.

referred to as *vortexes*. They are the whirlpools or eddies in which the field has its seat or origin. A turbulent field is sometimes also called a *vortex field*.

An arbitrary vector field can exhibit both turbulent and nonturbulent characteristics, caused by the simultaneous presence of sources and vortexes. The turbulent and nonturbulent components of this field are, however, linearly independent. In other words, an arbitrary vector field

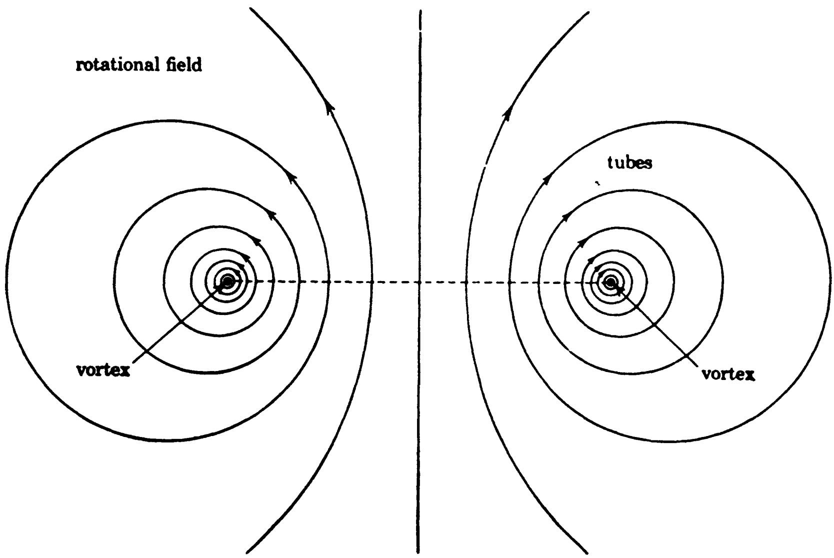


FIG. 2. A vector field which is solely turbulent.

may always be represented as the linear superposition of two independent components, one of which is a purely turbulent and the other a purely potential* field.

These matters, together with a number of useful vector operations and their interrelations as well as the geometrical interpretations of them, are discussed in the following articles. One of the various operations encountered here is the linear transformation discussed in Ch. III. The coefficients of this transformation (elements of its matrix) may be functions of the space co-ordinates, so that for each point in space, a particular vector transform is associated with any given vector. This

*This rather common designation for a nonturbulent field is appropriate because a vector function representing the gradient of a scalar potential is inherently nonturbulent, as is shown in detail subsequently.

transformation function is called a *tensor** *of valence 2*. The order of the tensor is that of its matrix. For ordinary space, therefore, the tensor is of the third order. The coefficients of the transformation are referred to as the *components* of the tensor.

A tensor of higher valence is a function which, at every point in space, associates a tensor of the next lower valence with any given vector. A

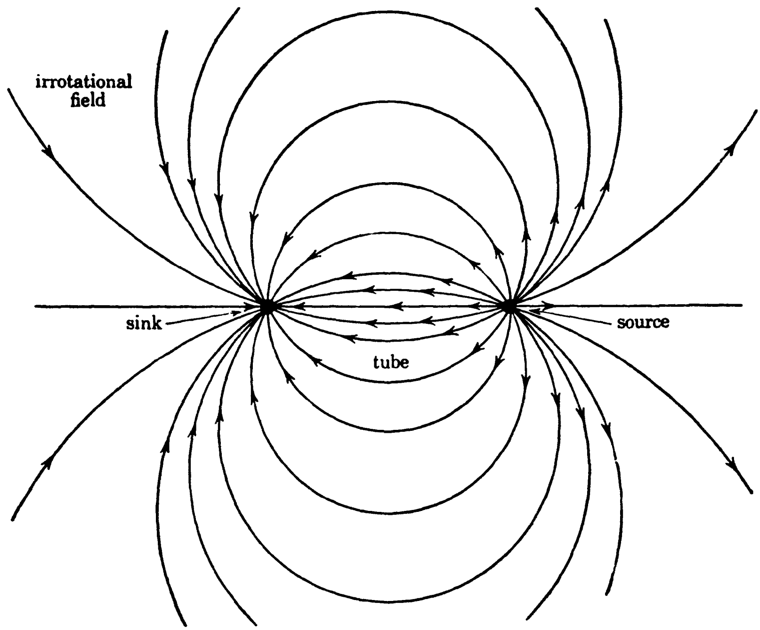


FIG. 3. A nonturbulent field.

tensor of order n and valence v has n^v components. For example, a tensor of the third order (for ordinary space) and valence 3, has 3^3 or 27 components. Its matrix may be regarded as a three-dimensional array. In this classification, a vector is sometimes referred to as a tensor of valence 1,† and a scalar as a tensor of valence 0.

*The name *tensor* originated when this kind of function was first used in connection with problems dealing with stresses in elastic media.

†In this interpretation of a vector, its components are regarded as defining a row matrix, and the linear transformation (tensor) corresponding to this matrix is a single linear equation like one of the equations of the set 3, Ch. III. This transformation is seen to transform a given vector x into a scalar (single component), but the classification of a vector as a tensor of valence 1 nevertheless appears to be inaccurate because it confuses the vector with a transformation. It would be more proper to say that the *components* of a vector (not the vector itself) may be regarded as those of a tensor of valence 1.

Since the analysis of a physical problem usually requires a co-ordinate system of some kind, the various vector operations discussed in this chapter are expressed not only in vector form but also in terms of an assumed system of co-ordinates. In most of the detailed formulations, the rectangular Cartesian system, being the simplest, is chosen. Unless mention is made to the contrary, a right-hand system of axes is assumed. The latter is so named because a right-hand screw, turning in the direction of the shortest route from the positive x-axis to the positive y-axis, advances in the direction of the positive z-axis. Transformations of the important vector operations to some of the orthogonal curvilinear co-ordinate systems more frequently encountered in practical problems are given in a subsequent article.

2. The scalar product

The *scalar* product of two vectors A and B (also called the *inner* product) is defined as the product of their magnitudes multiplied by the cosine of the angle between them. The scalar product is denoted by a dot placed between the symbols A and B; hence

$$A \cdot B = |A|\,|B| \cos \theta \qquad [1]$$

in which θ is the angle included between the two vectors.*

According to this definition, the scalar product may alternatively be regarded as the product of the length of either vector with the projection of the other upon it. If one of the vectors, for example, B, is given by the vector sum (according to the parallelogram law of addition) of two other vectors, as in the expression

$$B = C + D \qquad [2]$$

then

$$A \cdot B = A \cdot (C + D) = A \cdot C + A \cdot D \qquad [3]$$

This result is seen to be true because the projection of B upon A is evidently equal to the sum of the projections of C and D upon A. Hence the distributive law holds for scalar products.

When it is necessary to express the scalar product in terms of the components of the vectors A and B with reference to a rectangular Cartesian co-ordinate system, it is convenient to define a set of unit vectors having the directions of the x-, y-, and z-axes. These unit vectors

*According to this notation, which is attributed to Gibbs, the scalar product is sometimes also referred to as the *dot product*. An alternative notation quite frequently found in the literature is to indicate the scalar product by enclosing the symbols for the two vectors in parentheses, thus: $A \cdot B = (AB)$.

are denoted respectively by the letters i, j, and k. The components of the vectors — that is, their projections upon the x-, y-, and z-axes — are denoted respectively by A_x, A_y, A_z, and B_x, B_y, B_z.

The vectors themselves may then be written

$$A = iA_x + jA_y + kA_z \qquad [4]$$

and

$$B = iB_x + jB_y + kB_z$$

The terms in these equations are vector components, and the right-hand sides represent vector sums. The scalar product of A and B may now be written

$$A \cdot B = (iA_x + jA_y + kA_z) \cdot (iB_x + jB_y + kB_z) \qquad [5]$$

Since the distributive law holds, the right-hand side of Eq. 5 may be replaced by the sum of nine component scalar products such as $(iA_x) \cdot (iB_x)$, $(iA_x) \cdot (jB_y)$, etc. The unit vectors i, j, k are mutually at right angles to each other. Hence the scalar product of any one of these with any other one is zero, and the scalar product of any one with itself is unity; that is

$$\begin{aligned} i \cdot j = j \cdot k = k \cdot i = 0 \\ i \cdot i = j \cdot j = k \cdot k = 1 \end{aligned} \qquad [6]$$

Consequently, only three of the nine component scalar products resulting from Eq. 5 have nonzero values, so that

$$A \cdot B = A_xB_x + A_yB_y + A_zB_z \qquad [7]$$

The scalar product is thus given by the sum of the products of the corresponding components of the two vectors. This result is true only for a rectangular co-ordinate system (discussed in Art. 4, Ch. III) and no longer holds when the co-ordinate axes make oblique angles with each other.

It is clear from the definition of the scalar product that the commutative law holds; that is,

$$A \cdot B = B \cdot A \qquad [8]$$

Since the scalar product of two vectors yields a scalar, a triple scalar product such as might be denoted by $A \cdot B \cdot C$ has no meaning, and the question whether the associative law holds hence does not arise.

3. The vector product

The vector product of two vectors A and B (also called the outer product) is defined as a vector whose magnitude is given by the product

of the magnitudes of A and B, multiplied by the sine of the angle between them. The direction of the vector product is related to the directions of A and B by the right-hand screw rule in the sense that a right-hand screw, turning in the direction of the shortest route from the tip of A to the tip of B (assuming that these emanate from a common point*), advances in the direction of the vector product.

In connection with this definition, it is significant to observe that the angle θ between A and B, which enters into the determination of the magnitude of the vector product, is that angle through which the right-hand screw defining the direction of the resulting vector is turned in passing from the tip of A to the tip of B. This is usually taken to be the smaller of the two supplemental angles through which it is possible to turn in passing from one to the other of any two coterminous vectors A and B. The result, however, is the same if the larger of these two angles is chosen, because the direction is then reversed and the algebraic sign of the magnitude is reversed also.

In the Gibbs notation, the vector product is denoted by a cross, placed between the symbols for the two vectors.† Thus the magnitude of the vector product is expressed by

$$|A \times B| = |A|\ |B| \sin\theta \qquad [9]$$

This expression is recognized geometrically as being numerically equal to the area of the parallelogram defined by the two coterminous vectors A and B.

If the cosines of the angles between the normal to the surface of this parallelogram (in the direction of the vector $A \times B$) and the x-, y-, and z-axes of a rectangular co-ordinate system are denoted respectively by $\cos(n,x)$, $\cos(n,y)$, and $\cos(n,z)$, then

$$\begin{aligned} V_x &= |A \times B| \cos(n,x) \\ V_y &= |A \times B| \cos(n,y) \\ V_z &= |A \times B| \cos(n,z) \end{aligned} \qquad [10]$$

represent the projections of the area of the parallelogram upon the yz, zx, and xy planes respectively. Since

$$\cos^2(n,x) + \cos^2(n,y) + \cos^2(n,z) = 1 \qquad [11]$$

*In other words, the vectors are for the moment assumed to be coterminous, as they usually are. The definition of the vector product (and the scalar product also) is, however, independent of whether the vectors are coterminous. If they are not, then the direction of the vector product may perhaps be more easily visualized by first displacing one of the vectors parallel to itself until the two become coterminous.

†For this reason, the vector product is sometimes also referred to as the *cross-product*. An alternative method for indicating the vector product is to enclose the symbols for the two given vectors in square brackets, thus: $A \times B = [AB]$.

it follows that

$$|A \times B| = \sqrt{V_x^2 + V_y^2 + V_z^2} \qquad [12]$$

and hence V_x, V_y, and V_z are seen to be the components of the vector product

$$V = A \times B \qquad [13]$$

In other words, the component areas given by Eqs. 10 are recognized as having the properties of vector components, and their vector resultant is identified with the vector product as defined above.

Observe in this connection that a component area (projection of the surface of the parallelogram determined by the vectors A and B upon one of the three co-ordinate planes), as defined by one of the three Eqs. 10, is not merely a geometrical projection, but in addition involves an algebraic sign which reverses if the direction of the normal is reversed (replacing the angles of the cosine functions by their supplements). That is, the algebraic signs of the components 10 are controlled by the right-hand screw rule for the vector product.

If the projections of the vectors A and B upon the yz, zx, and xy planes are denoted by $A^{(yz)}$, $B^{(yz)}$, and so forth, it follows from these considerations that the vector components of V are given by

$$\begin{aligned} iV_x &= A^{(yz)} \times B^{(yz)} \\ jV_y &= A^{(zx)} \times B^{(zx)} \\ kV_z &= A^{(xy)} \times B^{(xy)} \end{aligned} \qquad [14]$$

The magnitude of one of these components, such as V_x, for example, is given by

$$V_x = |A^{(yz)}|\ |B^{(yz)}| \sin \theta_{yz} \qquad [15]$$

in which θ_{yz} is the angle included between $A^{(yz)}$ and $B^{(yz)}$. Replacing this angle by the difference between the angles which these vectors separately make with the y-axis, applying the trigonometric identity for the sine of the difference between two angles, and noting that the components of $A^{(yz)}$ and $B^{(yz)}$ are those of A and B on the Y and Z axes, it is found that

$$V_x = A_yB_z - A_zB_y \qquad [16]$$

and similarly that

$$V_y = A_zB_x - A_xB_z \qquad [17]$$

and

$$V_z = A_xB_y - A_yB_x \qquad [18]$$

These are the components of the vector product expressed in terms of

those of the two given vectors A and B. If the vector B is the resultant of two other vectors, that is, if

$$B = C + D \tag{19}$$

the decomposition of C and D into their rectangular components, and substitution into Eqs. 16, 17, and 18 show that

$$V = A \times B = A \times (C + D) = A \times C + A \times D \tag{20}$$

Hence it follows that the distributive law holds with regard to the vector product.

Conversely, if the distributive law is assumed to hold, the vector product

$$A \times B = (iA_x + jA_y + kA_z) \times (iB_x + jB_y + kB_z) \tag{21}$$

may be replaced by the sum of nine component vector products. According to the definition of the vector product,

$$i \times i = j \times j = k \times k = 0 \tag{22}$$

and

$$\begin{aligned} i \times j &= k = -j \times i \\ j \times k &= i = -k \times j \\ k \times i &= j = -i \times k \end{aligned} \tag{23}$$

so that three of the nine terms represented by Eq. 21 become zero, and the remaining six yield

$$A \times B = i(A_yB_z - A_zB_y) + j(A_zB_x - A_xB_z) + k(A_xB_y - A_yB_x) \tag{24}$$

This result is seen to agree with that stated by Eqs. 16, 17, and 18.

It is useful to recognize that the vector product may be written in the following determinant form:

$$A \times B = \begin{vmatrix} i & j & k \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \tag{25}$$

The Laplace expansion of this determinant in terms of the elements of its first row yields the vector product in the form given by Eq. 24.

Although the distributive law holds for the vector product, the commutative law evidently does not, since

$$A \times B = -B \times A \tag{26}$$

The triple vector product

$$A \times B \times C \tag{27}$$

has a unique meaning only when the order in which the products are to

be carried out is indicated. Thus in the association indicated by

$$A \times [B \times C] \tag{28}$$

the product $B \times C$ is formed first, and then the vector product of A with this vector is determined. The result is normal to $B \times C$ and hence is a vector which lies in the plane determined by the vectors B and C. In the association indicated by

$$[A \times B] \times C \tag{29}$$

the product $[A \times B]$ is formed first, and then the product of this vector with the vector C is determined. The result in this case must be normal to $A \times B$, and hence is a vector which lies in the plane determined by the vectors A and B.

Evidently

$$A \times [B \times C] \neq [A \times B] \times C \tag{30}$$

so that the associative law does not hold for multiple vector products.

Since the resultant vector for the triple product 28 lies in the plane of the vectors B and C, it must be possible to express this vector as a linear combination of B and C, that is,

$$A \times [B \times C] = \beta B + \gamma C \tag{31}$$

An evaluation according to the form given by Eq. 24 for the vector product shows that

$$A \times [B \times C] = (A \cdot C)B - (A \cdot B)C \tag{32}$$

which agrees with Eq. 31, in which

$$\beta = (A \cdot C) \qquad \text{and} \qquad \gamma = -(A \cdot B) \tag{33}$$

The triple product 29, on the other hand, is

$$[A \times B] \times C = -C \times [A \times B] = C \times [B \times A] \tag{34}$$

This has the form of the triple product in Eq. 32 with A and C interchanged. Hence

$$[A \times B] \times C = (A \cdot C)B - (B \cdot C)A \tag{35}$$

which is a vector lying in the plane determined by the vectors A and B, as stated above.

4. The scalar triple product

The following combination of a vector and a scalar product

$$A \cdot [B \times C] \tag{36}$$

in which A, B, and C are arbitrary vectors, is called a ***scalar triple product.***

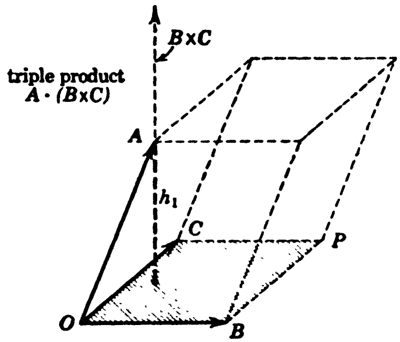
The definition of the scalar product in the form given by Eq. 7, together with the determinant form 25 for the vector product, shows that the scalar triple product may be expressed as the value of the determinant

$$A \cdot [B \times C] = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} \qquad [37]$$

The result is, of course, a scalar.

Since the transpose of a determinant has the same value,* the rows in the determinant 37 may be written alternatively as columns. The value reverses its algebraic sign when any pair of rows are interchanged, but it remains unchanged if this interchanging is done twice in succession. Since the cyclic order of the letters A, B, C can be changed to B, C, A by two interchanges, and to C, A, B by two more interchanges, it follows that

$$A \cdot [B \times C] = B \cdot [C \times A] = C \cdot [A \times B] \qquad [38]$$



h_1 = projection of vector A upon $B \times C$

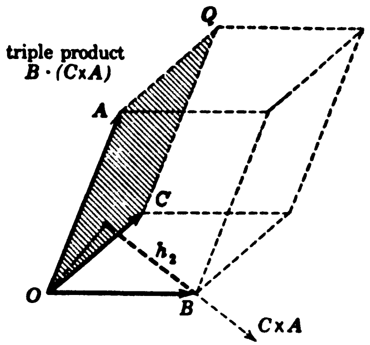
$|B \times C|$ = area of parallelogram $OBPC$

FIG. 4. A geometrical interpretation of a scalar triple product.

A geometrical interpretation for the scalar triple product is readily given as shown in Fig. 4. The magnitude of $[B \times C]$ equals the area of the parallelogram determined by B and C; its direction is normal to the plane of this parallelogram. If the scalar product is interpreted as the length of the vector $[B \times C]$ multiplied by the projection of A upon it, the scalar triple product is seen to be equal to the product of the area of the parallelogram determined by B and C, multiplied by the component of A normal to this surface. The result evidently represents the volume of the parallelepiped of which three coterminous edges coincide in length and

*See Art. 2, Ch. I. Specifically, this is there stated as the property VIII.

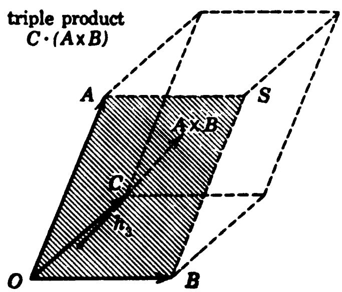
direction with the three vectors A, B, and C. The parallelogram determined by B and C is regarded as the base of this parallelepiped, and the normal component of A is its altitude. In the equivalent forms given by Eq. 38, the base of the parallelepiped is alternatively regarded as defined by the vectors C and A, or A and B. These alternate interpretations are illustrated in Figs. 5 and 6. Each of the three expressions in Eq. 38 represents the volume of the same parallelepiped.



h_2 = projection of vector B upon $C \times A$

$|C \times A|$ = area of parallelogram $OCQA$

FIG. 5. Alternate interpretation of a scalar triple product.



h_3 = projection of vector C upon $A \times B$

$|A \times B|$ = area of parallelogram $OBSA$

FIG. 6. Alternate interpretation of a scalar triple product.

In connection with this scalar triple product, observe that the brackets enclosing the vector product of B and C in the expression 36 are significant because they indicate the order in which the operations are to be carried out. Thus, if this triple product were written $A \cdot B \times C$, it might be thought that $A \cdot B$ could be carried out first if desired. This procedure, however, yields a scalar, and the subsequent vector product with C then has no meaning.

5. The gradient

A single-valued scalar function of the space co-ordinates x, y, z is denoted by the symbol U. It is a function of position or location only. The points in space at which U has a given value, for example, C, define a surface which is referred to as a *constant-value* surface. Any number of such surfaces, for various assumed values of the constant C, may be mapped. In particular, it is expedient to map a series of constant-value surfaces for values of C which differ by integer multiples of some chosen interval. A familiar example of such a map is that for the two-dimensional altitude function in a geographical terrain. Here the function U is constant

along lines instead of surfaces. These are called *contour lines*, and the resulting plot is spoken of as a *contour map*.

Such a map places the variation of the function U in evidence, since the function evidently changes slowly in those regions where the lines or surfaces are far apart, and rapidly where they are closely spaced. The rate at which U varies in any given direction at a point in space is determined approximately by the ratio which the interval chosen for the constant C-values has to the distance measured between two neighboring surfaces in the given direction at the point in question. If the interval in the C-values at this point is allowed to become smaller and smaller, the corresponding limiting value of the ratio accurately yields the desired rate of change of U. This value is called the *directional derivative* of U.

It is apparent that the directional derivative of U is a maximum at a given point if the derivative is taken in a direction normal to the constant-value surface passing through that point, because the distance between neighboring surfaces is evidently smallest in the normal direction. This maximum value of the directional derivative is called the *normal derivative* of U.

As a function of the space co-ordinates, the normal derivative of U appears to have the properties of a vector function. The truth of this statement may be seen from the fact that if dn is the differential distance in the direction of the normal between two neighboring constant-value surfaces for which C differs by dC, and if ds is the distance between these surfaces in any other direction s, then, except for differentials of higher order,

$$dn = ds \cos \theta \tag{39}$$

where θ is the angle between the normal and the direction s. It follows that

$$\frac{\partial U}{\partial s} = \frac{\partial U}{\partial n}\frac{dn}{ds} = \frac{\partial U}{\partial n}\cos \theta \tag{40}$$

in which $\partial U/\partial n$ is the normal derivative of U.

This result shows that if the normal derivative is regarded as a vector pointing in the direction of the normal, the derivative of U in any direction s is given by the projection of this vector upon a line having that direction. The normal derivative, therefore, has true vector character. This vector is called the *gradient* of U at the point at which the normal derivative is evaluated.

The gradient is defined as pointing in that direction in which U increases and it is evidently a function of the space co-ordinates, since its magnitude and direction depend upon the point at which $\partial U/\partial n$ is evaluated. For a geographical altitude function, the gradient at any

point indicates the direction of steepest ascent, and its magnitude equals the maximum rate of change of altitude with distance at that point.

In equations, the gradient is written in the abbreviated form: *grad U*. If n denotes distance measured along the normal at any point in the direction in which U increases, and n_1 represents a unit vector in this direction, the gradient is expressed by the vector equation

$$\text{grad } U = n_1 \frac{\partial U}{\partial n} \qquad [41]$$

If s_1 denotes a unit vector in any direction s, the component of the gradient in that direction is given, with the help of the scalar product and Eq. 40, by

$$\text{grad}_s\, U = s_1 \cdot \text{grad } U = s_1 \cdot n_1 \frac{\partial U}{\partial n} = \frac{\partial U}{\partial n} \cos \theta = \frac{\partial U}{\partial s} \qquad [42]$$

In a rectangular co-ordinate system, the components of the gradient are

$$\text{grad}_x\, U = \frac{\partial U}{\partial x} \qquad \text{grad}_y\, U = \frac{\partial U}{\partial y} \qquad \text{grad}_z\, U = \frac{\partial U}{\partial z} \qquad [43]$$

and hence

$$\text{grad } U = i \frac{\partial U}{\partial x} + j \frac{\partial U}{\partial y} + k \frac{\partial U}{\partial z} \qquad [44]$$

A more compact form for this expression is obtained by defining the so-called Hamiltonian operator*

$$\nabla = i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z} \qquad [45]$$

The equivalent of Eq. 44 then reads

$$\text{grad } U = \nabla U \qquad [46]$$

The operator ∇ may in some respects be formally treated as a vector with the components $\partial/\partial x$, $\partial/\partial y$, $\partial/\partial z$. These, however, cannot in general be manipulated as though they were ordinary algebraic coefficients. They are differential parameters (or operators) of the first order, and hence perform the operation of differentiation upon whatever function follows them.

For example, if W is another scalar function, the rule for differentiating

*Since the symbol for this operator is an inverted Greek capital delta, it is frequently referred to by the name " del," and grad U is alternatively called " del of U." Another name for the operator ∇ is " nabla," after the Greek name for a harp, which this symbol resembles in form.

a product shows that

$$\nabla(UW) = U\nabla W + W\nabla U \tag{47}$$

or

$$\text{grad }(UW) = U \text{ grad } W + W \text{ grad } U \tag{48}$$

Another example is the scalar triple product 36 in which the operator ∇ takes the place of the vector A. Here the relation 38 is not applicable without due attention to the implied operation of differentiation in the multiplication of ∇ with the functions B and C. The correct evaluation of the expression $\nabla \cdot [B \times C]$ is given in a subsequent article (see Art. 16) after the operations $\nabla \cdot A$ and $\nabla \times A$ have been discussed. The present remarks are made merely to caution the reader against any careless manipulation of the vector operator ∇.

An important property of the gradient comes to light from a consideration of the so-called *line integral* of this vector function evaluated for an arbitrary path extending between any two points in space. Thus if

$$f = \text{grad } U \tag{49}$$

then the integral

$$\int_a^b f \cdot \mathbf{ds} \tag{50}$$

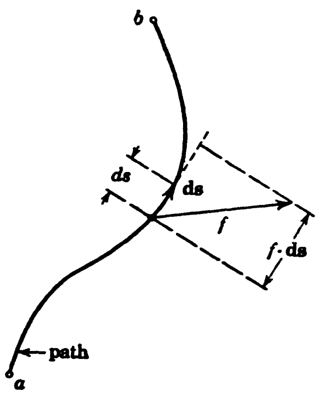


FIG. 7. A scalar product involving the differential vector **ds**.

is referred to as the line integral of the vector function f between the points $a(x_1,y_1,z_1)$ and $b(x_2,y_2,z_2)$. These two points are assumed to be connected by a continuous path or curve S of arbitrary form. Distance along this curve is denoted by the symbol s. The differential *vector* distance **ds** is at any point tangential to the path in the direction of continuous progress along it from a to b, and in magnitude equals the scalar increment of length ds. The scalar product $f \cdot \mathbf{ds}$ then equals the component of f coincident with the direction of travel along the path at any intermediate point, multiplied by the corresponding path increment. This relationship is illustrated in Fig. 7.

If f represents the force on some particle which is constrained to follow this path (like a bead on a bent wire), in the absence of friction the line integral 50 evidently yields the total work done by the force as the particle travels along the path from a to b.

According to Eqs. 40 and 41, it is recognized that

$$f \cdot \mathbf{ds} = n_1 \cdot \mathbf{ds} \frac{\partial U}{\partial n} = \frac{\partial U}{\partial s} ds = dU \qquad [51]$$

in which dU is the differential increment of work done by the force f in moving the particle through the path increment ds.* Since U is a function of position only, dU is a total differential. Hence

$$\int_a^b f \cdot \mathbf{ds} = \int_a^b \text{grad } U \cdot \mathbf{ds} = \int_a^b dU = U(x_2,y_2,z_2) - U(x_1,y_1,z_1) \qquad [52]$$

From this result it is concluded that *the line integral of the gradient between any two points is independent of the path joining these points.* Hence if the particle is subsequently returned to the point a along any other path from b to a, the total work done by the force f as the particle traverses the closed circuit from a to b and back to a, is zero.

Symbolically, the line integral extending over a closed circuit is indicated by a circle placed upon the integral sign. Thus the vector function defined as the gradient has the property that

$$\oint \text{grad } U \cdot \mathbf{ds} = 0 \qquad [53]$$

It is important to observe that the truth of the results expressed by Eqs. 52 and 53 depends upon the single valuedness of the function U. This restriction on the function U, which is stated in the opening paragraph of the present article, is not always met by functions dealt with in practical problems. Further considerations, necessary when U is multivalued, are given in Art. 14 of this chapter.

In the above argument, the force f is regarded as due to some external agency which is causing the motion of the particle along a given path. This external or driving force must, of course, be balanced by an equal but oppositely directed force of reaction. The latter is due to the inherent properties of the medium or system through which the particle is moved. If F denotes this force of reaction, evidently

$$F = -\text{grad } U \qquad [54]$$

In this connection, U is spoken of as the *potential* function of the system, and F is the vector field of force associated with U. If U_0 denotes the value of U at some chosen datum point, then $U - U_0$ represents the

*One may alternatively verify this conclusion by writing $f = i\frac{\partial U}{\partial x} + j\frac{\partial U}{\partial y} + k\frac{\partial U}{\partial z}$, and $\mathbf{ds} = i\,dx + j\,dy + k\,dz$, whence, according to Eq. 7,

$$f \cdot \mathbf{ds} = \frac{\partial U}{\partial x} dx + \frac{\partial U}{\partial y} dy + \frac{\partial U}{\partial z} dz = dU$$

work which must be done upon the particle to move it from the datum to the point to which U refers. This work is called the ***potential energy*** of the system consisting of the given particle and the medium in which it is embedded.

A simple example is the potential energy of a particle (usually considered to have unit mass) located at some altitude above sea level. The function F is then referred to as the earth's gravitational field of force. Here the significance of the relation 53 is readily visualized. Thus, whatever work may be done by the external force f while the particle traverses part of the closed path is returned to the external agency during the traversal of the remainder of the circuit. This would be the case, for example, if one were to carry some object around a closed path on the side of a mountain.

The work which the external agency may contribute during the traversal of a certain portion of the path is thought of as stored by the force field through which the particle is moved, and the significance of Eq. 53 is that such stored energy is not lost but may be completely regained. A force field which has this property of conserving whatever energy increment may be imparted to it is called a *conservative* field. This is the property of a vector field defined by the gradient function.

Such a vector field is evidently irrotational in character, for if the flow map for this field were assumed to contain any lines which close upon themselves, one of these could be chosen as the path for the integral in Eq. 53, and the value of this integral would then certainly not be zero. Hence the gradient always defines a purely potential field, all the lines of which must emanate from sources and terminate upon sinks. A conservative field is always irrotational.

Conversely, if for a given vector field F it is known that

$$\oint F \cdot \mathbf{ds} = 0 \qquad [55]$$

it must be possible to define a potential function U such that F is given by Eq. 54.

6. The divergence

The sources of a potential field are sometimes thought of as concentrated at points or distributed along filaments or over surfaces. Whereas such source distributions may, in the discussion of certain physical problems, be convenient from an analytic point of view, they nevertheless are idealizations which require a proper mathematical interpretation. For the present discussion, the sources of a potential field are considered to be continuously distributed throughout space.

According to the hydrodynamic analogy, a region in which sources are located is one from which fluid emanates in a continuously distributed fashion, like oil seeping up through the pores of a bed of quicksand. (The pores as well as the rates of flow through them are to be thought of as being infinitesimal.) The object of the present discussion is to formulate some means for describing the fluid productivity of an infinitesimal element of space in a source-filled region. In other words, some measure of source density or intensity is needed in order that a given distribution of sources may be described and the relation of this distribution to the associated field intensity may be determined.

The net rate at which fluid emanates from a small but finite productive region may be measured through integrating the rate of flow over a surface enclosing this region. If this net rate is divided by the enclosed volume, the resulting figure represents an *average* rate of productivity per unit volume for this region. The actual rate of productivity per unit volume may, of course, vary from point to point throughout the region. At any given point, it is expressible as the limit of the average rate obtained through shrinking the enclosure about that point until the contained volume becomes infinitesimal.

This limiting value is a convenient measure of the intensity of the source region at any point, and it is referred to as the *divergence* of the flow field at that point. Although the flow field is a vector function, its divergence is clearly a scalar.

If A is any vector function, then in accordance with the discussion just given, its divergence (abbreviated *div A*) may be mathematically defined by

$$\operatorname{div} A = \underset{\int dv \to 0}{\text{limit}} \left\{ \frac{\displaystyle\oint_{\text{closed surface}} A \cdot \mathbf{da}}{\displaystyle\int_{\text{enclosed volume}} dv} \right\} \qquad [56]$$

The circle on the integral sign in the numerator indicates that the integration extends over a *closed* surface. In this integral, **da** is a *vector* surface increment. Its magnitude equals the scalar differential area *da* at any point on the surface, and its direction is that of the outwardly directed normal at that point. The scalar product $A \cdot \mathbf{da}$, therefore, represents the product of the normal component of A and the scalar differential area *da* at the same point. This equals the rate at which fluid passes through the surface element *da*, if A is thought of as representing the velocity field of a fluid.

The integral in the numerator of Eq. 56 is then seen to equal the total rate at which fluid passes outward through the closed surface. The integral

in the denominator of this expression represents the enclosed volume. The resulting limit of the ratio of these two integrals yields the divergence of A at the point about which the closed surface is shrunk by the limiting process. Its value is in general different for different points in space.

In order for div A to be calculated when the vector function A is given, Eq. 56 must be further evaluated. For this purpose, the enclosed

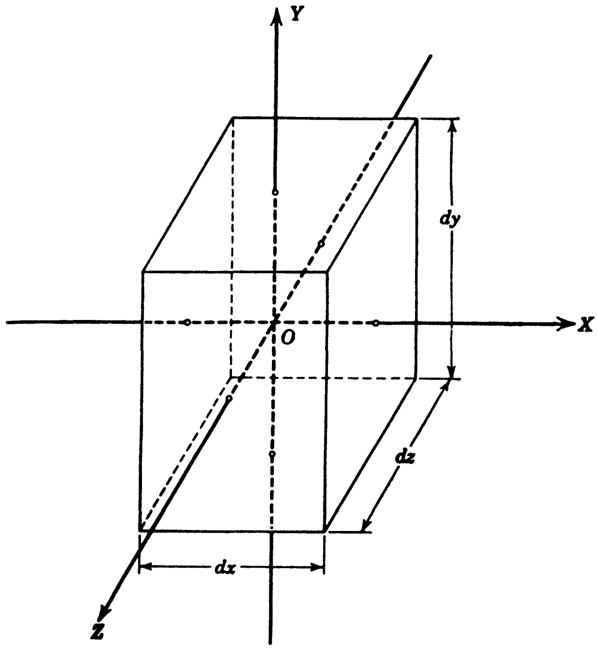


FIG. 8. Enclosed volume assumed in the calculation of div A.

volume is effectively assumed in the form of a rectangular parallelepiped with its center located at the origin of a rectangular co-ordinate system, and three of its coterminous edges coincident in direction with the co-ordinate axes as shown in Fig. 8. If the sides of this parallelepiped are identified with the differentials dx, dy, dz, the desired result is at once obtained without the necessity of subsequently carrying out the limiting process.

The vector function A is assumed to be finite and continuous in the vicinity of the parallelepiped, so that, except for differentials of the second order, the variation of A is linear throughout this infinitesimal region. If A_x, A_y, A_z denote the values of the components of A at the

origin, for the surface of the parallelepiped

$$\oint A \cdot d\mathbf{a} = \left(A_x + \frac{\partial A_x}{\partial x}\frac{dx}{2}\right) dy\, dz - \left(A_x - \frac{\partial A_x}{\partial x}\frac{dx}{2}\right) dy\, dz$$
$$+ \left(A_y + \frac{\partial A_y}{\partial y}\frac{dy}{2}\right) dz\, dx - \left(A_y - \frac{\partial A_y}{\partial y}\frac{dy}{2}\right) dz\, dx \qquad [57]$$
$$+ \left(A_z + \frac{\partial A_z}{\partial z}\frac{dz}{2}\right) dx\, dy - \left(A_z - \frac{\partial A_z}{\partial z}\frac{dz}{2}\right) dx\, dy$$

which yields

$$\oint A \cdot d\mathbf{a} = \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) dx\, dy\, dz \qquad [58]$$

The enclosed volume is given by

$$\int dv = dv = dx\, dy\, dz \qquad [59]$$

Hence, in rectangular co-ordinates, Eq. 56 evaluates to

$$\operatorname{div} A = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \qquad [60]$$

By means of the Hamiltonian operator defined by Eq. 45, and the form for the scalar product given by Eq. 7, the result stated in Eq. 60 may be written

$$\operatorname{div} A = \nabla \cdot A \qquad [61]$$

It is significant that the partial derivatives in Eq. 60 are always to be evaluated at the same point in space at which the divergence of A is desired. If the result is numerically positive, the point in question is a source; if it is negative, the point is a sink. If the result is zero over a finite region, the latter is source free. On the other hand, if div A is zero throughout *all* space, it may be concluded either that A is zero everywhere or that this vector function describes a field which is purely rotational in character.* The possibility that A might be constant throughout all space must be discarded on the physical ground that any field must vanish at infinity (exceptions to this rule are due to idealizations which are physically unrealizable).

7. Gauss's law

According to the definition of the divergence as expressed by Eq. 56,

*This alternative is considered in greater detail in Art. 10.

it follows that

$$\oint_{\text{closed surface}} A \cdot \mathbf{da} = \int_{\text{enclosed volume}} \operatorname{div} A \, dv \qquad [62]$$

Since the integrand in the right-hand integral, namely, div $A\,dv$, represents the fluid productivity for the volume element dv, whence this integral yields the total productivity for the finite region, a result that is alternately given through integrating the fluid flow over the enclosing surface.

This result, which is known as *Gauss's law*, formally represents the transformation of a surface integral into a volume integral, or vice versa.

It should be observed that the hydrodynamic analogy, used in the preceding article to lend concreteness to the definition of the divergence, tacitly implies that the "fluid" be *incompressible*. An arbitrary vector function may be likened to the flow of an incompressible or ideal fluid. In this light, Gauss's law becomes almost self-evident, since it states merely that all the incompressible fluid produced within a given region (this is the volume integral of div A) must issue from the enclosing surface.

Gauss's law is generally applicable to any vector function for which the divergence exists, that is, for which the partial derivatives given in Eq. 60 can be formed. If this condition is not, the vector function A is in general regular and continuous throughout the region involved. When these conditions are not fulfilled, the desired transformation may in certain cases still be achieved by means of special manipulations to which the following discussion is pertinent.

8. Idealized source distributions

If the region contains points at which A is infinite or discontinuous, the divergence at such points cannot be evaluated by means of the formula given in Eq. 60. Such situations occur in practical problems under idealized assumptions. For example, it is sometimes convenient to assume that a source region is two-dimensional or one-dimensional, or even that it has zero dimensions. These idealized source distributions are referred to respectively as a *surface* distribution, a *filamental* distribution, or as a *point* source.

For the continuous distribution of sources considered in the previous article, the source density or productivity per unit volume may conveniently be denoted by some symbol such as ρ, and defined in conformance with its analogy to electric charge density by the equation

$$\operatorname{div} A = \rho \qquad [63]$$

Equation 62 may then be written

$$\oint_{\text{closed surface}} A \cdot \mathbf{da} = \int_{\text{enclosed volume}} \rho \, dv \qquad [64]$$

Here the volume integral on the right represents the total productivity of the region enclosed by the surface over which the integral on the left extends. Equation 64 evidently holds regardless of how the source density ρ is distributed throughout the enclosed region, and hence it is possible to assume (if convenient) that the total productivity is concentrated at one or more points. These are then referred to as point sources.

In certain physical problems, the actual distribution of source density approximates this idealization closely enough to justify such an assumption, and thus a simplification in the resulting mathematical relationships for the determination of the field is made available.

A filamental source density distribution is similarly an idealization found convenient in certain types of physical problems. It is significant, however, that for the point or filamental types of source distributions the concept of the divergence does not apply. External to the points or filaments, the divergence is, of course, zero, whereas for points coinciding with these idealized sources the divergence as defined above becomes infinite.

For a point source, the total productivity (point charge) may be defined by the relation

$$e = \rho \, dv \qquad [65]$$

in which the density ρ is, of course, infinite because e is assumed finite. Similarly, for the filamental distribution, a productivity per unit length may be defined as

$$q = \rho \, da \qquad [66]$$

the cross-section of the filament being denoted by da. Here again, ρ must be considered infinite.

Actually there can be no infinite density ρ, but the practical examples to which the idealizations expressed by Eqs. 65 and 66 apply are such that the geometrical relations are closely approximated when the actual finite source region or filament cross-section is replaced by the infinitesimals dv and da respectively.

In an analogous manner, a surface distribution of sources is defined as having a productivity per unit area given by

$$\sigma = \rho \, ds \qquad [67]$$

in which the differential thickness ds represents the actual small finite thickness of a continuous distribution in the form of a layer.

For this surface distribution, it is useful to extend the definition of the divergence by making use of the relations 63 and 64. A differential surface element da of the source layer is assumed to have a thickness ds which is vanishingly small in comparison with the surface dimensions of da. This is a perfectly admissible assumption which in fact adds to the preciseness of the definition 67. The integral on the left of Eq. 64, applied to the surface enclosing this element, reduces to the flow outward from the two opposite faces of the layer element, so that this equation yields

$$\sigma\, da = (A_{n1} + A_{n2})\, da \tag{68}$$

in which A_{n1} and A_{n2} are the outwardly directed normal components of A on the two sides of the layer. By analogy to Eq. 63 it is then possible to define a so-called *surface divergence* per unit area, given by

$$\mathrm{div}_\sigma A = \sigma = A_{n1} + A_{n2} \tag{69}$$

9. The scalar potential function associated with a given source distribution

If A is a vector function describing a potential field, it is related to its sources by Eq. 63, and to a scalar potential function U by an equation similar to Eq. 54. Hence

$$\mathrm{div}\, A = -\,\mathrm{div}\,\mathrm{grad}\, U = \rho \tag{70}$$

and it follows that a potential function U may be associated with any given source distribution ρ.

By means of Eqs. 46 and 61, this relationship may be written

$$\nabla \cdot \nabla U = -\rho \tag{71}$$

and, with the help of Eqs. 44 and 60, the interpretation of this form is readily seen to be expressed by

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = -\rho \tag{72}$$

Alternatively, the scalar product $\nabla \cdot \nabla$ may be interpreted as the resultant operator

$$\nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{73}$$

which yields Eq. 72 when applied to the potential function U.

The differential operator of the second order defined by Eq. 73 is referred to as the *Laplacian* operator, and the form

$$\nabla \cdot \nabla U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} \tag{74}$$

is spoken of as the "Laplacian of U." For purposes of abbreviation, $\nabla \cdot \nabla$ is alternatively denoted by the symbols ∇^2 and Δ, so that Eqs. 71 or 72 are frequently written either as

$$\nabla^2 U = -\rho \qquad [75]$$

or as

$$\Delta U = -\rho \qquad [76]$$

the latter notation being the simpler but subject to some objection on the ground that the symbol Δ might, according to its more common interpretation, be confused with the notation for an increment.

The significant point of the present discussion is the fact that a given source distribution defines an associated scalar potential function by means of Eq. 72, and this in turn determines the associated vector function A by means of the relation

$$A = -\operatorname{grad} U \qquad [77]$$

This view of the relation between the functions A and ρ may seem to be more roundabout than the apparently simple one expressed by Eq. 63, but in the solution of many practical problems it proves to be a more convenient method of determining A from ρ, because Eq. 72 involves scalar functions only and hence is frequently more easily integrated than Eq. 63. Once the function U is obtained, it is usually not difficult to determine A from Eq. 77 because doing so involves differentiation only.

Moreover, there are certain problems in which the source distribution ρ is not known, but, instead, U and its normal derivatives are known over boundaries of given geometrical form to which the unknown source distribution is restricted. For the solution of these so-called *boundary-value problems*, the attack by means of Eq. 72 is the only possible one. At all points not located on the boundary, $\rho = 0$, and this equation reads

$$\nabla^2 U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0 \qquad [78]$$

A function which formally satisfies this differential equation, and in addition meets the stated boundary conditions imposed upon U, constitutes the desired solution.

Equation 72 is known as *Poisson's* equation, and the corresponding homogeneous equation 78 as *Laplace's*. The latter equation is of importance in connection with any problem in which the mapping of flow fields is essential, whether this is accomplished by analytic or by graphical means.*

*An application of this method to the determination of electric circuit parameters is given in *Electric Circuits*, Ch. I, Art. 6a, pp. 16 ff.

10. The curl of a turbulent* vector field

In the present article, attention is turned from the potential field to what might be called its complement — the turbulent or rotational field. Here the flow lines form closed circuits, so that the line integral of the vector function A taken around a closed path is in general not zero. That is, for a rotational vector function A

$$\oint A \cdot \mathbf{ds} \neq 0 \qquad [79]$$

The seat or origin of a turbulent field lies in its vortexes or whirlpools. It is clear that the greater the intensity of the whirlpools in a given region, the greater is the value of the integral 79 evaluated for various contours within this region. It is, therefore, reasonable to suggest that the value of this integral be used as a basis for defining the vortex density at any point.

If the vector A is thought of as representing the force acting upon a particle, the integral 79 represents the work done by this force field as the particle is allowed to traverse a closed path. The path may, for the moment, be thought of as circular, with a radius r. The value of the integral 79 divided by 2π then represents an average torque with respect to a concentric axis normal to the plane of this circular path. This average torque evidently varies with the angular orientation of the axis for a fixed location of the center of the circular path. There will evidently be one orientation for which the average torque is a maximum. As the radius is allowed to shrink until it becomes zero, the maximum average torque becomes zero also, but the ratio of this torque to the area of the circle is found to approach a finite limit when the force A has finite values. It is this limiting value of the ratio of the maximum average torque to the enclosed area which thus proves to be a useful measure of the turbulence of the field A at any given point. This value, except for the factor 2π, is called the curl of A at that point.

It is not essential for the definition of the curl that the closed path be circular. It is essential, however, that this path lie in a plane so that the orientation of the normal be clearly defined, for the curl is a vector having this same orientation. Thus the curl of A is given in magnitude

*In some more recent considerations of hydrodynamic fields, the term *turbulent* is used to designate an entirely random character that specifically does not permit a mathematical representation in terms of either the gradient or the curl. Consideration of fields having such random character is not included in the present discussion. The term *turbulent* in this volume is used merely as a means of distinguishing the rotational from the irrotational character of a field.

by the maximum value of

$$\lim_{\int da \to 0} \left[\frac{\displaystyle\oint_{\text{plane contour}} A \cdot \mathbf{ds}}{\displaystyle\int_{\text{enclosed surface}} da} \right] \tag{80}$$

as the plane of the contour assumes all possible orientations. In direction, the curl of A coincides with the normal to the plane of the contour,

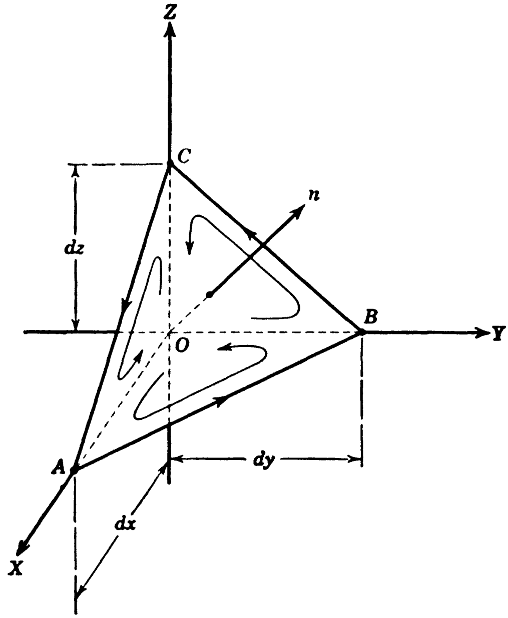


FIG. 9. An infinitesimal tetrahedron to illustrate the vector character of the curl.

pointing in agreement with the advance of a right-hand screw which turns so as to correspond to the traversal of the contour in the evaluation of the closed line integral. The vector, curl A, is thus uniquely defined in direction as well as in magnitude.

Just as the divergence of a potential field is at any point a measure of the density of its sources at that point, so the curl of a turbulent field establishes an analogous relationship between the intensity of that field and its vortex density at any point.

The vector character of the curl may be demonstrated by means of the geometrical configuration in Fig. 9, showing an infinitesimal triangular

contour ABC whose normal n has an arbitrary orientation relative to a reference system of co-ordinates. The sides of this triangle together with portions of the co-ordinate axes form the six edges of a tetrahedron with vertexes O, A, B, C. The sides OBC, OCA, OAB, normal respectively to the X-, Y-, and Z-axes, have infinitesimal areas which are denoted by da_x, da_y, and da_z. The area of the side ABC opposite the vertex O is denoted by da. Since da_x, da_y, da_z, are the projections of da upon the co-ordinate planes YZ, ZX, and XY respectively, one may write

$$\frac{1}{da} = \frac{\cos(n,x)}{da_x} = \frac{\cos(n,y)}{da_y} = \frac{\cos(n,z)}{da_z} \qquad [81]$$

in which $\cos(n,x)$, $\cos(n,y)$, $\cos(n,z)$ are the direction cosines of the normal n.

It is likewise clear from the geometry of the configuration that

$$\oint_{ABC} A \cdot \mathbf{ds} = \oint_{OBC} A \cdot \mathbf{ds} + \oint_{OCA} A \cdot \mathbf{ds} + \oint_{OAB} A \cdot \mathbf{ds} \qquad [82]$$

inasmuch as the net integration represented by the right-hand side of this equation involves traversals in both directions along the edges of the tetrahedron emanating from the vertex O (see the circulatory arrows in Fig. 9). In terms of the notation

$$R = \frac{\oint_{ABC} A \cdot \mathbf{ds}}{da} \qquad [83]$$

and

$$R_x = \frac{\oint_{OBC} A \cdot \mathbf{ds}}{da_x} \qquad R_y = \frac{\oint_{OCA} A \cdot \mathbf{ds}}{da_y} \qquad R_z = \frac{\oint_{OAB} A \cdot \mathbf{ds}}{da_z} \qquad [84]$$

the results expressed by Eqs. 81 and 82 show that

$$R = R_x \cos(n,x) + R_y \cos(n,y) + R_z \cos(n,z) \qquad [85]$$

Subject to the condition

$$\cos^2(n,x) + \cos^2(n,y) + \cos^2(n,z) = 1 \qquad [86]$$

the expression on the right-hand side of Eq. 85 attains the maximum value

$$R_{max} = \sqrt{R_x^2 + R_y^2 + R_z^2} \qquad [87]$$

for the particular orientation of the normal given by*

$$\cos(n,x) = \frac{R_x}{R_{max}} \qquad \cos(n,y) = \frac{R_y}{R_{max}} \qquad \cos(n,z) = \frac{R_z}{R_{max}} \qquad [88]$$

Since the relations expressed by the last two equations are the familiar ones existing between a vector and its rectangular components, the desired proof is completed.

It is clear that the curl is an axial vector similar to that representing a mechanical torque, the right-hand screw rule serving in both instances to relate the rotational motion to the axial direction of the vector. The components of this vector with respect to a rectangular co-ordinate system may be determined by noting that the component of the curl in any direction s is given by

$$\operatorname{curl}_s A = \lim_{\int da \to 0} \left[\frac{\displaystyle\oint_{\text{plane contour normal to } s} A \cdot ds}{\displaystyle\int_{\text{enclosed surface}} da} \right] \qquad [89]$$

in which the direction of s is linked with the direction of traversal in the line integral by means of the right-hand screw rule.

The x-component of the curl is thus found from consideration of a closed contour lying in the yz-plane. For convenience, this is taken to be the contour of a differential rectangle with its center at the origin, as

---

*To maximize $f(x,y,z) = ax + by + cz$ subject to the condition $x^2 + y^2 + z^2 = 1$, form: $f = ax + by + cz - \frac{\lambda}{2}(x^2 + y^2 + z^2 - 1)$. Equating the partial derivatives to zero gives:

$$\frac{\partial f}{\partial x} = a - \lambda x = 0 \qquad x = \frac{a}{\lambda}$$

$$\frac{\partial f}{\partial y} = b - \lambda y = 0 \qquad y = \frac{b}{\lambda}$$

$$\frac{\partial f}{\partial z} = c - \lambda z = 0 \qquad z = \frac{c}{\lambda}$$

The condition then yields $\lambda = \sqrt{a^2 + b^2 + c^2}$ so that one has

$$f_{max} = \sqrt{a^2 + b^2 + c^2}$$

and the particular values of the variables yielding this maximum are

$$x = \frac{a}{f_{max}} \qquad y = \frac{b}{f_{max}} \qquad z = \frac{c}{f_{max}}$$

For other applications of the method of determining conditioned maxima see Arts. 4 and 5, Ch. IV.

illustrated in Fig. 10. Since the co-ordinate axes form a right-hand system, the X-axis points upward from the plane of the paper, and the direction of traversal of the rectangular contour is as indicated by the arrows. It is clear that only the y- and z-components of A contribute to the x-component of the curl.

The components of A in the direction of traversal along the sides 1–2, 2–3, 3–4, and 4–1 may be denoted respectively by A_{12}, A_{23}, A_{34}, and A_{41}. If A_y and A_z denote the values of the y- and z-components of A at the origin, then, except for differential contributions of higher order (assuming A to be finite and continuous throughout the region covered by the rectangle),

$$\begin{aligned} A_{12} &= A_z + \frac{\partial A_z}{\partial y}\frac{dy}{2} \\ A_{23} &= -A_y - \frac{\partial A_y}{\partial z}\frac{dz}{2} \\ A_{34} &= -A_z + \frac{\partial A_z}{\partial y}\frac{dy}{2} \\ A_{41} &= A_y - \frac{\partial A_y}{\partial z}\frac{dz}{2} \end{aligned} \qquad [90]$$

and hence

$$\begin{aligned} \oint A \cdot \mathbf{ds} &= A_{12}\,dz + A_{23}\,dy + A_{34}\,dz + A_{41}\,dy \\ &= \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) dy\,dz \end{aligned} \qquad [91]$$

According to Eq. 89, the x-component of curl A, therefore, becomes

$$\text{curl}_x A = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \qquad [92]$$

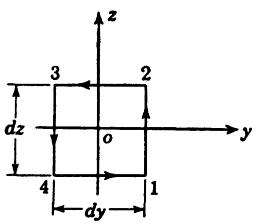


FIG. 10. A differential rectangle centered at the origin.

In an analogous fashion, or through simply advancing the cyclic order of the letters x, y, z, the y- and z-components of curl A are found to be

$$\text{curl}_y A = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \qquad [93]$$

$$\text{curl}_z A = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \qquad [94]$$

Comparing the turbulent with the potential field again, it is interesting to observe the significance of the conditions under which

$$\oint A \cdot \mathbf{ds} = 0 \qquad [95]$$

According to the discussion in Art. 5, these conditions are stated by

$$A = -\operatorname{grad} U \tag{96}$$

or

$$A_x = -\frac{\partial U}{\partial x} \qquad A_y = -\frac{\partial U}{\partial y} \qquad A_z = -\frac{\partial U}{\partial z} \tag{97}$$

Now since

$$\frac{\partial^2 U}{\partial x\,\partial y} = \frac{\partial^2 U}{\partial y\,\partial x}, \text{ etc.} \tag{98}$$

it follows that the conditions 97 may alternatively be stated in the form

$$\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} = 0 \tag{99}$$

$$\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} = 0 \tag{100}$$

$$\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} = 0 \tag{101}$$

The expressions on the left-hand sides of these equations are the components of curl A given in Eqs. 92, 93, and 94. Hence it is seen that

$$\operatorname{curl} A = 0 \tag{102}$$

becomes the necessary and sufficient condition for the vanishing of the line integral as expressed by Eq. 95.

If over a given region the curl of A vanishes at all points, the line integral of A for any closed path within this region vanishes also,* and A may there be represented as the gradient of a scalar potential function as expressed by Eq. 96. Conversely, if the vector function A is the gradient of a scalar, its curl must be zero; that is, A must be nonturbulent. *The gradient then, has no curl.* In symbols,

$$\operatorname{curl} \operatorname{grad} U \equiv 0 \tag{103}$$

This statement is verified through substituting the relations 97 into the Eqs. 99, 100, and 101.

It is useful to observe that the components of curl A given by Eqs. 92, 93, and 94 may be combined into a single compact vector expression by means of the determinant form

$$\operatorname{curl} A = \begin{vmatrix} i & j & k \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} \tag{104}$$

*This statement is subject only to the restrictions pointed out in Art. 14, which need not be considered at the moment.

The Laplace development of this determinant for the elements of its first row is readily seen to yield

$$i \operatorname{curl}_x A + j \operatorname{curl}_y A + k \operatorname{curl}_z A \qquad [105]$$

in which the respective components are the expressions 92, 93, and 94.

By recalling the determinant form 25 for the vector product and the definition of the Hamiltonian operator ∇ as given by Eq. 45, it is recognized that an alternative compact form for curl A reads

$$\operatorname{curl} A = \nabla \times A \qquad [106]$$

In summary, the three important vector operations, gradient, divergence, and curl, are seen to be expressible in terms of the Hamiltonian operator, thus

$$\begin{aligned} \operatorname{grad} U &= \nabla U \\ \operatorname{div} A &= \nabla \cdot A \\ \operatorname{curl} A &= \nabla \times A \end{aligned} \qquad [107]$$

11. Stokes's law

According to the definition of the curl as expressed by Eq. 80 or in component form by Eq. 89, it follows that

$$\underset{\text{any closed contour}}{\oint A \cdot \mathbf{ds}} = \underset{\substack{\text{any surface bounded} \\ \text{by that contour}}}{\int (\operatorname{curl} A) \cdot \mathbf{da}} \qquad [108]$$

This result, which is known as Stokes's law, formally represents the transformation of a closed line integral into a surface integral, or vice versa. Stokes's law is the counterpart of Gauss's law in the sense that it expresses for the purely rotational field what Gauss's law expresses with reference to the potential field.

It should be observed that the closed contour in Eq. 108 is not restricted to lie in a plane, and that the surface bounded by this contour may have any shape. For example, the contour may be visualized as a warped hoop and the surface as that of a rubber membrane bounded by the hoop but allowed to be stretched out sideways into any form whatever. If the hoop is thought of as being moderate in size, one may consider blowing the rubber membrane out sideways until it becomes inflated like a balloon, which may bulge backward over the ring, etc.

The vector surface increment **da**, which is normal to the surface, points in that direction which is determined from the direction of traversal of the boundary by the right-hand screw rule. This correlation is shown in Fig. 11. In Fig. 12 the surface is for convenience drawn as a plane. The

scalar product (curl A) · **da** equals the line integral of A for the closed boundary of any one of the surface elements. The surface integral represents the sum of all these elemental closed line integrals. This sum

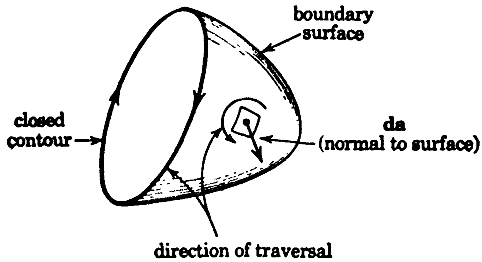


FIG. 11. Correlation of contour traversal with element traversal when bounded surface does not lie in a plane.

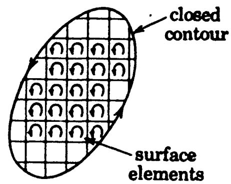


FIG. 12. The closed line integral equals the sum of the elemental line integrals.

evidently equals the line integral around the large boundary because the common boundaries of the surface elements are traversed in both directions and their contribution to the surface integral is zero. Thus the validity of Stokes's law, as stated by Eq. 108, is established.

12. THE VORTEX DISTRIBUTION OF A TURBULENT FIELD

For the turbulent field, it is convenient to define a *vortex density* J by means of the relation

$$\text{curl } A = J \tag{109}$$

in which J is analogous to the current density occurring in the study of electricity and magnetism.* Stokes's law, Eq. 108, then yields

$$\oint_{\text{closed contour}} A \cdot \mathbf{ds} = \int_{\text{enclosed surface}} J \cdot \mathbf{da} \tag{110}$$

Equations 109 and 110 are analogous to Eqs. 63 and 64. They relate the turbulent field to its vortex distribution J just as the potential field is related to its source distribution ρ. The vortex density J is the cause of the turbulent field just as the source density ρ is the cause of a potential field.

The essential difference is that ρ is a scalar function of the space coordinates whereas J is a vector function. In general, J is a finite and

*In this analogy the vector function A represents the magnetic field intensity and should not be confused with the magnetic vector potential which is customarily denoted by the letter A.

continuous function; that is, the vortexes are thought of as continuously distributed over certain regions just as the source density ρ is in general continuously distributed. In certain practical problems, however, it is convenient to idealize the vortex distribution. The situation is again analogous to the idealized source distributions discussed in Art. 8, except that there is no vortex analogue of the point source.*

There is, however, a vortex analogue of the filamental source distribution. It is spoken of as a *vortex thread*. The *amount* I of the vortex thread is defined by the relation

$$I = J\,da \tag{111}$$

in which da is written in place of the actual finite cross-section of the filament. Since I is finite and da differential, this idealization requires the concept of an infinite vortex density J. The latter, as well as the vector I, is everywhere directed tangentially to the filament. The vortex thread may be thought of as a filament along which are concentrated whirlpools of infinite intensity, and about which a fluid is set into a swirling motion. The central portion of a tornado or "twister," and that of a "waterspout," are hydrodynamic examples permitting an idealized representation by means of a vortex thread.

A surface distribution of vortexes may be thought of as the result of many threads placed side by side like the warp in a weaver's loom. The vector moment I has the same value and direction for all the threads. A surface density g may here be defined by the relation

$$g = J\,ds \tag{112}$$

in which ds represents the thickness of the surface or layer. There is no whirling action around the individual threads, that is, through the surface, because of the cancellation of this action along adjacent sides of the threads. The vortex surface merely causes a translatory motion of fluid in opposite directions along its two sides. Thus the tangential component of the vector field A is observed to change suddenly as the point of observation is shifted from one side of the vortex surface to the other at any given point.

Figure 13 shows the vortex surface in cross-section, the cut being parallel to the tangential components of A and at right angles to the direction of g. The line integral in Eq. 110 is evaluated for the closed rectangular path with the differential sides ds and $d\ell$. Here ds is assumed to be so small compared to $d\ell$ that the corresponding contributions to the line integral are negligible in comparison with those for the sides $d\ell$.

*In this connection it may be observed that one does have a vortex analogue of the double point source (the doublet or dipole), namely, the *vortex ring*.

Equation 110 then yields

$$(A_{t1} - A_{t2})\, d\ell = J\, ds\, d\ell = g\, d\ell \qquad [113]$$

in which A_{t1} and A_{t2} are the tangential components of A on the two sides of the surface.

By analogy to Eq. 109, it is then possible to define a *surface curl* per unit length (in the $d\ell$ direction), given by

$$\text{curl}_\sigma A = g = A_{t1} - A_{t2} \qquad [114]$$

This vector points upward from the surface of the paper (the assumed direction of the vortex density g) in Fig. 13. If n_{12} denotes a unit vector normal to the vortex surface pointing from side 1 to side 2, the direction of the surface curl is contained in the right-hand screw rule for the vector product in the definition

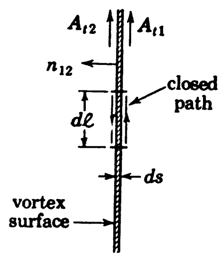


FIG. 13. A vortex surface in cross section.

$$\text{curl}_\sigma A = g = (A_{t1} - A_{t2}) \times n_{12} \qquad [115]$$

13. THE VECTOR POTENTIAL FUNCTION ASSOCIATED WITH A GIVEN VORTEX DISTRIBUTION

The present article seeks to show that the vortex distribution of a turbulent field determines a vector potential function in a manner similar, except for certain anti-parallelisms, to that relating a scalar potential function to the source distribution of a potential field, as discussed in Art. 9.

The first step in this argument is to observe that if Stokes's law, Eq. 108, is applied to a *closed* surface, the line integral on the left-hand side of this equation is zero. This fact may be visualized through first considering the surface to be balloon shaped and the closed contour or boundary to be the small loop located at the throat through which the balloon is inflated. As this loop is contracted until the balloon is finally tied off, the closed contour shrinks to zero and the line integral vanishes.

According to Gauss's law, Eq. 62, it follows, therefore, that

$$\oint_{\text{closed surface}} (\text{curl}\, A) \cdot da = \int_{\text{enclosed volume}} \text{div}\, (\text{curl}\, A)\, dv = 0 \qquad [116]$$

Since the size of the enclosed volume is arbitrary, it may be considered to be infinitesimal; whence this result yields

$$\text{div curl}\, A \equiv 0 \qquad [117]$$

The conclusion is that the vector field given by the function curl A is *source-free;* in other words, *curl A* is a purely turbulent vector field. This conclusion is true for any vector function A. The relation 117 is an identity.

The curl then, may be said to have no divergence. This statement is the complement of the one made in Art. 10 to the effect that the gradient has no curl. Whereas the gradient function represents a pure potential field, the curl represents a purely turbulent field.

The identity 117 may readily be checked independently through substituting the components of the curl given by Eqs. 92, 93, and 94 for the components A_x, A_y, A_z in the expression for the divergence as given by Eq. 60.

If a given vector function B is known to represent a purely turbulent field, that is, if

$$\operatorname{div} B = 0 \tag{118}$$

the result just obtained permits the function B to be expressed as

$$B = \operatorname{curl} A \tag{119}$$

the thought being that the vector function A is thereby specified in terms of B, or vice versa. Here it must be observed, however, that whereas B is uniquely determined in terms of A by means of Eq. 119, the reverse is not true.

This operation is readily appreciated from the fact that if the vector function A is assumed to be perfectly general, it may represent a field which exhibits *both* potential as well as turbulent characteristics. In other words, A may be represented by

$$A = T + P \tag{120}$$

in which T is the source-free (or turbulent) component of A, and P is its curl-free (or potential) component, that is

$$\operatorname{div} T = 0 \tag{121}$$

and

$$\operatorname{curl} P = 0 \tag{122}$$

Substituting the expression 120 for A into Eq. 119 yields

$$B = \operatorname{curl} T \tag{123}$$

because of the relation 122.

It thus becomes clear that Eq. 119 alone does not determine the whole function A, but only its source-free component. If the vector function A is introduced by Eq. 119 merely for convenience, that is, as an auxiliary function in the course of an analysis, this difficulty may be overcome by

the further demand that

$$\operatorname{div} A = 0 \qquad [124]$$

The vector function A is then uniquely characterized in terms of B by means of the *two* relations 119 and 124.

This situation is similar to having given a purely potential field F, and assuming that a scalar potential function U may thereby be defined in terms of the relation

$$F = -\operatorname{grad} U \qquad [125]$$

In this case, U is determined only within an additive arbitrary *constant*, since the gradient of the latter is zero. Just so, Eq. 119 determines A only within an additive arbitrary vector function representing a potential field, since the curl of the latter is zero. Setting the constant component of U equal to some chosen value (for example, zero) is analogous to setting the divergence of A equal to zero in the present discussion.*

It may now be supposed that the given source-free field represented by B has a vortex distribution described by the vector function J; that is

$$\operatorname{curl} B = J \qquad [126]$$

From Eq. 119 it then follows that a vector function A is related to this vortex distribution by means of the equation

$$\operatorname{curl}\operatorname{curl} A = J \qquad [127]$$

subject to the condition expressed by Eq. 124.

The function A which is thus related to a given vortex distribution J by means of Eqs. 124 and 127 is called the *vector potential* associated with B, because it plays the same role relative to B and its vortexes that the scalar potential U plays relative to the associated nonturbulent field F and its sources. If the vector potential A is found from a given vortex distribution J by means of Eqs. 124 and 127 the turbulent field B is determined from Eq. 119. This method of determining B from J, although apparently roundabout, is frequently found to be more convenient than that of attempting a solution directly in terms of Eq. 126.

It is necessary to interpret further the repeated curl operation in Eq. 127, as is readily done by means of Eqs. 92, 93, and 94 for the components of the curl. Thus the x-component of curl curl A is seen to be given by

$$\frac{\partial}{\partial y}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) =$$

$$\frac{\partial^2 A_y}{\partial x\,\partial y} - \frac{\partial^2 A_x}{\partial y^2} - \frac{\partial^2 A_x}{\partial z^2} + \frac{\partial^2 A_z}{\partial x\,\partial z} \qquad [128]$$

*Any other disposition of the value of div A, appropriate to the circumstances pertaining to a specific problem, may likewise be made.

Adding and subtracting from this expression the term $\frac{\partial^2 A_x}{\partial x^2}$, and appropriately grouping the result, show that

$$\text{curl}_x\,(\text{curl}\,A) = \frac{\partial}{\partial x}\left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) - \left(\frac{\partial^2 A_x}{\partial x^2} + \frac{\partial^2 A_x}{\partial y^2} + \frac{\partial^2 A_x}{\partial z^2}\right) \quad [129]$$

By means of Eqs. 60 and 74, this may be written

$$\text{curl}_x\,(\text{curl}\,A) = \frac{\partial}{\partial x}\,(\text{div}\,A) - \nabla^2 A_x \quad [130]$$

in which ∇^2 is the Laplacian operator defined by Eq. 73.

The y- and z-components of curl curl A are found in a like manner, or more simply through advancing the cyclic order of the variables x, y, z in Eq. 129. The result is

$$\text{curl curl}\,A = \left(i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z}\right)(\text{div}\,A) - \nabla^2(iA_x + jA_y + kA_z) \quad [131]$$

which is

$$\text{curl curl}\,A = \text{grad div}\,A - \text{div grad}\,A \quad [132]$$

or

$$\text{curl curl}\,A = \nabla(\nabla\cdot A) - \nabla^2 A \quad [133]$$

It is interesting that although the gradient of the vector function A is not defined,* the Laplacian of A, or div grad A, is interpreted according to Eq. 131 as

$$\text{div grad}\,A = \nabla^2 A = i\,\nabla^2 A_x + j\,\nabla^2 A_y + k\,\nabla^2 A_z \quad [134]$$

Equation 127, subject to the condition 124, is now seen to yield the equation

$$\nabla^2 A = -J \quad [135]$$

relating the vector potential associated with B to its vortex distribution. The nicety of this result is that it is identical in form to Eq. 75 relating the scalar potential U to the source distribution ρ of a nonturbulent field. The only significant difference is that Eq. 135 involves vector functions whereas Eq. 75 involves scalar functions. This fact means that the vector equation 135 is equivalent to three scalar equations of like form, one for each of the three corresponding components of A and J.

*An interpretation of the gradient applicable also to vector functions is given in Art. 15 of this chapter.

14. The possibility of a multivalued potential function

It has been shown that if, throughout a given region, a vector field A is conservative so that curl A is identically zero at all points within this region, A may there be expressed as the gradient of a scalar potential function U. Furthermore, it has been pointed out that the existence of such a potential field is in general due to a distribution of sources which, according to the hydrodynamic analogy, are fluid-producing regions. One may be tempted to conclude, therefore, that a potential field can be due only to sources, and never to vortexes, because these latter are regions of turbulence and hence can produce only turbulent fields.

A moment's reflection, however, yields the thought that if a distribution of vortex density J is confined to a finite region of space, only within that region is the curl of the associated vector function A different from zero. Everywhere else the curl of A is zero, and the field is, therefore, potential (that is, conservative) in character.

A very important distinction, however, should be made between the conservative field which is due only to pure sources, and that which is due to vortex distributions restricted to certain excluded regions of space. To state the matter in another way, it is important to distinguish between a field which is conservative throughout *all* space and one which is conservative only within certain portions of space, or to the exclusion of certain portions. The field which is conservative everywhere can have its origin in true sources only, whereas the one which is conservative within reserved regions (finite or infinite in extent) can be due to vortexes as well as to sources, although the former are confined to lie outside the reserved regions.

This distinction is concerned primarily with the nature of the associated scalar potential function U. Thus, if the field is conservative throughout all space, or throughout a *simply connected* region δ (see Fig. 14), the associated potential function U there is *single valued;* but if the region over which the field is conservative is a *multiply connected* one (Fig. 15), the associated potential becomes a *multivalued* function.

A region is said to be simply connected if for *every* closed contour lying entirely within the region a surface bounded only by that contour can be constructed such that every point of the surface also lies within the region. The significance of this statement is best understood from an example of a region which is not simply connected. Thus the space outside a doughnut-shaped region is not simply connected because a closed contour which links the doughnut (passes through the hole) cannot form the sole boundary of a surface every point of which is required to lie outside the doughnut. If the doughnut is cut so that it no longer forms a closed ring, the space becomes simply connected. With the doughnut intact, the surrounding space is said to be *doubly* connected.

Another illustration of a doubly connected space is that surrounding a closed cylindrical surface of infinite extent like a straight tube which is infinitely long in both directions. Here a contour which encloses the cylinder cannot form the sole boundary of a surface which lies wholly

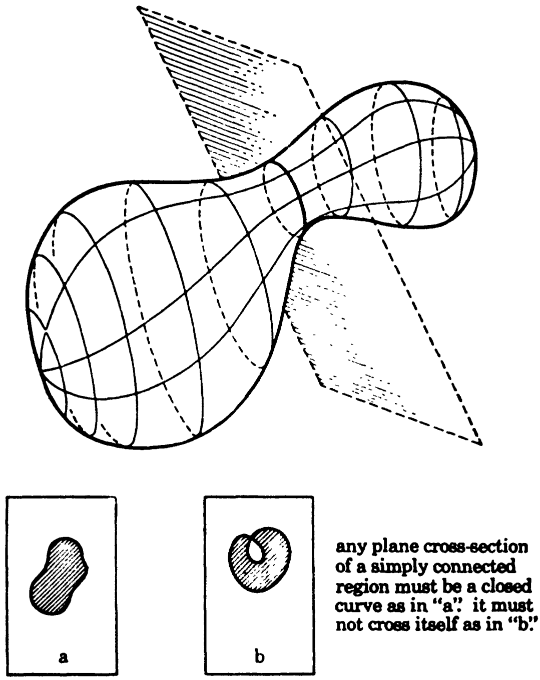


FIG. 14. A simply connected region.

outside the cylinder. If the space enclosed by the cylindrical surface is filled with vortexes whose cross-sectional distribution is given by the vector density J (directed longitudinally), then, since by assumption $J = 0$ outside the cylinder, the associated field function A satisfies the equations

$$\operatorname{curl} A = J \qquad [136]$$

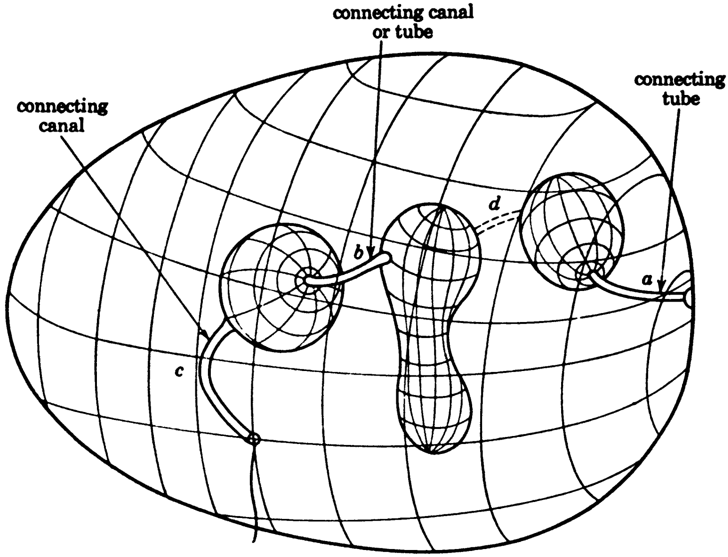
within the cylinder, and

$$\operatorname{curl} A = 0 \qquad [137]$$

outside the cylinder.

In the region outside the cylinder it is, therefore, possible to assume

$$A = -\operatorname{grad} U \qquad [138]$$



connecting tube d can be used instead of a but not both at the same time.
the connecting tubes, of infinitesimal cross-section, are the
so-called barriers of the region

FIG. 15. A multiply connected region.

With reference to Fig. 16, which shows a cross-section through the cylinder, the line integral of A formed for the closed path P_1 is evidently zero because it does not enclose the vortex region. This fact follows readily from Stokes's law. However, for the closed path P_2 the same law yields

$$\oint_{P_2} A \cdot \mathbf{ds} = \int J \cdot \mathbf{da} = I \qquad [139]$$

But by Eq. 138

$$\oint_{P_2} A \cdot \mathbf{ds} = -\oint_{P_2} (\text{grad } U) \cdot \mathbf{ds} \qquad [140]$$

and hence it follows that

$$\oint_{P_2} (\text{grad } U) \cdot \mathbf{ds} = -I \qquad [141]$$

If the line integral is thought of as extending from some point a on the path P_2 around the closed contour back to the same point a, this result

indicates that

$$\oint_{P_2} (\text{grad } U) \cdot \mathbf{ds} = U_a - U_a = I \qquad [142]$$

in which U_a is the value of the potential U at the point a. The only possible conclusion to be drawn from Eq. 142 is that the potential function U is not uniquely defined for points in the space surrounding the

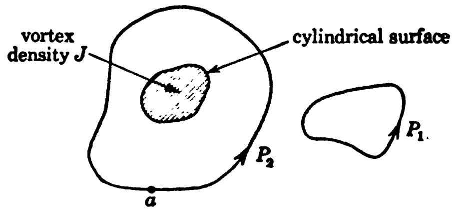


FIG. 16. One path of integration which encloses a vortex; another which does not.

region enclosed by the cylindrical surface. Indeed, if the line integral is extended *twice* around the contour P_2, the result is

$$U_a - U_a = 2I \qquad [143]$$

and for n circuitations

$$U_a - U_a = nI \qquad [144]$$

The potential U at any point such as a clearly is defined only within an additive integer multiple of I. This is the nature of the multivalued potential function U in this example.

The structure of the space surrounding the cylindrical region may be thought of as composed of leaves lying in the cross-sectional plane and winding about the cylinder as a winding stairway encircles a column. The density of these leaves longitudinally is infinite so that the process of encircling the cylinder any finite number of times around the contour P_2 without piercing any leaves involves no net longitudinal motion. Any point such as a is then referred to as $a_1, a_2, a_3, \cdots$ etc., according to an assumed numbering of these hypothetical leaves, and Eq. 144 may then be written more precisely in the form

$$U_{a_{k+n}} - U_{a_k} = nI \qquad [145]$$

Although the multivalued potential function U may, through such a mathematical artifice, be transformed into a single-valued one, the rather complicated interpretation of the space outside the cylindrical region is given here principally for the purpose of lending visual clarity to the

concept of multivaluedness. A practical problem involving these concepts arises in the subject of magnetostatics where the cylindrical region is identified with a current-carrying conductor, the density of the electric current being J, whereas the total conductor current has the value I.

15. The differentiation of scalar or vector functions with respect to the time

A given vector function A is considered to be a function of the space co-ordinates x, y, z and also of the time t, that is,

$$A = A\,(x,y,z,t) \qquad [146]$$

The behavior of this function is studied at some point P in space, having the co-ordinates x, y, and z. More specifically, this point is denoted by $P(x,y,z)$. This notation may refer to any specific point if proper values are assigned to x, y, and z. P is also referred to as "the point of observation."

In considering the time derivative of A at the point P, or the rate at which the vector A changes with respect to the time at the point of observation, two possibilities must be distinguished, according to whether the point of observation is stationary or is in motion relative to the reference co-ordinate system. In the former case, the variables x, y, and z are constant, and in the latter they are functions of the time — that is, the point P moves with a velocity v given by

$$v = iv_x + jv_y + kv_z \qquad [147]$$

with

$$v_x = \frac{dx}{dt} \qquad v_y = \frac{dy}{dt} \qquad v_z = \frac{dz}{dt} \qquad [148]$$

If the point P is stationary (x, y, z are assumed constant), the time derivative of A is understood to be the *partial* derivative $\partial A/\partial t$; if P is not stationary, the time derivative of A is given by the *total* derivative dA/dt.

Identical remarks apply to the time derivative of a scalar function U, which may, for example, be the potential energy of a particle located at the point P. If the particle is stationary, the potential energy may change with the time because of a time variation of the field in which it is embedded. This possibility is denoted by the partial derivative $\partial U/\partial t$. If the particle is in motion, the *net* rate of change of its potential energy with respect to the time is denoted by dU/dt, and is in general due to the combined effects of the time variation of the field in which the particle is embedded and of the particle's own motion through this field.

Neglecting differential effects of higher order than the first, the total

time derivative of the scalar function U is expressed by

$$\frac{dU}{dt} = \frac{\partial U}{\partial t} + \frac{\partial U}{\partial x}\frac{dx}{dt} + \frac{\partial U}{\partial y}\frac{dy}{dt} + \frac{\partial U}{\partial z}\frac{dz}{dt} \qquad [149]$$

The first term on the right-hand side of this equation represents the time variation of U on the assumption that the particle at the point P is stationary. The sum of the remaining three terms represents the time variation of U due solely to the motion of the particle through the field which (as far as these terms are concerned) is constant at any fixed point but varies from point to point. The net time variation dU/dt appears as a linear superposition of these two separate effects because the contribution of higher order derivatives may (according to the principles of the differential calculus) be neglected.

According to the definitions of the gradient and the scalar product, Eq. 149 may be written in the form

$$\frac{dU}{dt} = \frac{\partial U}{\partial t} + v \cdot \text{grad}\, U \qquad [150]$$

in which the velocity v of the point of observation is defined by Eqs. 147 and 148. If U is not an explicit function of the time — that is, if U is a function of the space co-ordinates alone — $\partial U/\partial t = 0$, and dU/dt is given by the second term in Eq. 150 alone. This term equals the component of grad U in the direction of v, multiplied by the magnitude of v. In other words this term equals the space rate of change of U in the direction of v, multiplied by the magnitude of v. It is convenient to regard this as a resultant operation upon the function U by defining the operator

$$(v \cdot \text{grad}) = (v \cdot \nabla) = v_x \frac{\partial}{\partial x} + v_y \frac{\partial}{\partial y} + v_z \frac{\partial}{\partial z} \qquad [151]$$

and by writing Eq. 150

$$\frac{dU}{dt} = \frac{\partial U}{\partial t} + (v \cdot \nabla)U \qquad [152]$$

The vector v in the operator defined by Eq. 151 may be replaced by any vector function B, and $(B \cdot \nabla)$ is then referred to as the "B-gradient." The symbolic operation $(B \cdot \nabla)U$ is read: "The B-gradient of U." Unlike the gradient of U, it is a scalar quantity.

The convenience resulting from defining this operator is seen when the operator is used in connection with a vector function A. Whereas the gradient of a vector function cannot be interpreted according to the definition of this vector operation, the B-gradient is readily seen to be applicable to vector functions, and yields a vector, thus

$$(B \cdot \nabla)A = i(B \cdot \nabla)A_x + j(B \cdot \nabla)A_y + k(B \cdot \nabla)A_z \qquad [153]$$

in which the x-component reads

$$(B \cdot \nabla)A_x = B \cdot \nabla A_x = B_x \frac{\partial A_x}{\partial x} + B_y \frac{\partial A_x}{\partial y} + B_z \frac{\partial A_x}{\partial z} \qquad [154]$$

Similar expressions apply to the y- and z-components.

The total variation of a vector function A with the time is, therefore, given by the expression

$$\frac{dA}{dt} = \frac{\partial A}{\partial t} + (v \cdot \nabla)A \qquad [155]$$

This is, in general, a function of the time. At any given instant, the vector $(v \cdot \nabla)A$ represents the space rate of change of A in the direction of v, multiplied by the magnitude of v. To state the case in another way, $(v \cdot \nabla)A$ is the vector change (increment) of A per unit displacement of the point of observation in the direction of v, multiplied by the magnitude of v.

A purely algebraic interpretation of the operation $(v \cdot \nabla)A$ is obtained by letting

$$(v \cdot \nabla)A = G = iG_x + jG_y + kG_z \qquad [156]$$

Then

$$\begin{aligned} \frac{\partial A_x}{\partial x} v_x + \frac{\partial A_x}{\partial y} v_y + \frac{\partial A_x}{\partial z} v_z &= G_x \\ \frac{\partial A_y}{\partial x} v_x + \frac{\partial A_y}{\partial y} v_y + \frac{\partial A_y}{\partial z} v_z &= G_y \\ \frac{\partial A_z}{\partial x} v_x + \frac{\partial A_z}{\partial y} v_y + \frac{\partial A_z}{\partial z} v_z &= G_z \end{aligned} \qquad [157]$$

The operation is thus recognized as amounting to a linear transformation of the vector v. If for abbreviation one writes

$$a_{xx} = \frac{\partial A_x}{\partial x}, \quad a_{xy} = \frac{\partial A_x}{\partial y}, \quad \cdots, \quad a_{zy} = \frac{\partial A_z}{\partial y}, \quad \cdots, \quad a_{zz} = \frac{\partial A_z}{\partial z} \qquad [158]$$

then the matrix of the transformation is

$$\mathfrak{A} = \begin{bmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{bmatrix} \qquad [159]$$

The operation $(v \cdot \nabla)A$ is thus regarded as a tensor of valence 2, which associates a vector G with a given vector v at any point in space. The elements of the matrix $\mathfrak{A}$, defined by Eq. 158, are the components of this tensor. Their values depend upon the vector function A. G is said to be a

linear vector function of v, the characteristics of this relationship being determined by the nature of A.

In passing it may be of interest to note that the operations of forming the curl or a vector product may also be regarded as linear vector transformations. Thus the vector, curl A, is expressible as a linear transformation of A with the skew-symmetrical operator matrix

$$\begin{bmatrix} 0 & -\dfrac{\partial}{\partial z} & \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} & 0 & -\dfrac{\partial}{\partial x} \\ -\dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial x} & 0 \end{bmatrix} \qquad [160]$$

and the vector product $A \times B$ is expressible as a linear transformation of B with the skew-symmetrical matrix

$$\begin{bmatrix} 0 & -A_z & A_y \\ A_z & 0 & -A_x \\ -A_y & A_x & 0 \end{bmatrix} \qquad [161]$$

The time derivative of the product of two vector functions is evaluated according to the familiar rule for the differentiation of a product. For the scalar product, the result reads

$$\frac{d}{dt}(A \cdot B) = \frac{dA}{dt} \cdot B + A \cdot \frac{dB}{dt} \qquad [162]$$

and similarly for the vector product

$$\frac{d}{dt}[A \times B] = \frac{dA}{dt} \times B + A \times \frac{dB}{dt} \qquad [163]$$

in which it is important to preserve the order of the two functions.

16. Additional useful vector relations

Although the discussions of the preceding articles suffice to determine the results of more complicated operations involving combinations of scalar and vector functions, it is useful to have available several additional formulas covering those relationships which are encountered more frequently in practical problems. The following items are concerned with derivations of this sort.

(i) *The divergence of the product of a scalar and a vector.*

The scalar function is denoted by U and the vector function by A.

Then

$$\text{div }(UA) = \frac{\partial}{\partial x}(UA_x) + \frac{\partial}{\partial y}(UA_y) + \frac{\partial}{\partial y}(UA_z) \qquad [164]$$

By the rule for the differentiation of a product, this is seen to yield

$$\text{div }(UA) = U \text{ div } A + A \cdot \text{grad } U \qquad [165]$$

(ii) *The curl of the product of a scalar and a vector.*

Here the rule for the differentiation of a product may be expressed in the following way:

$$\text{curl }(UA) = \nabla \times (UA) = \nabla \times (UA)_U + \nabla \times (UA)_A \qquad [166]$$

in which the subscripts mean that the functions U or A are to be treated as constants. According to the definition of the curl by Eq. 104 it is readily seen that

$$\nabla \times (UA)_U = U \text{ curl } A \qquad [167]$$

and

$$\nabla \times (UA)_A = -A \times \text{grad } U \qquad [168]$$

Hence

$$\text{curl }(UA) = U \text{ curl } A - A \times \text{grad } U \qquad [169]$$

(iii) *The divergence of a vector product.*

This may be written

$$\text{div }[A \times B] = \nabla \cdot [A \times B] = \nabla \cdot [A \times B]_A + \nabla \cdot [A \times B]_B \qquad [170]$$

the subscripts having the same significance as before. From Eqs. 26 and 38,

$$\nabla \cdot [A \times B]_A = -A \cdot [\nabla \times B] = -A \cdot \text{curl } B \qquad [171]$$

$$\nabla \cdot [A \times B]_B = B \cdot [\nabla \times A] = B \cdot \text{curl } A \qquad [172]$$

so that

$$\text{div }[A \times B] = B \cdot \text{curl } A - A \cdot \text{curl } B \qquad [173]$$

(iv) *The curl of a vector product.*

Here

$$\text{curl }[A \times B] = \nabla \times [A \times B] = \nabla \times [A \times B]_A + \nabla \times [A \times B]_B \qquad [174]$$

Applying Eq. 32 for the triple vector product gives

$$\nabla \times [A \times B]_A = A(\nabla \cdot B) - (A \cdot \nabla)B \qquad [175]$$

and

$$\nabla \times [A \times B]_B = (B \cdot \nabla)A - B(\nabla \cdot A) \qquad [176]$$

The operations A-gradient and B-gradient are defined by Eq. 151 in the previous article. Thus

$$\operatorname{curl}\,[A \times B] = A \operatorname{div} B - B \operatorname{div} A + (B \cdot \nabla)A - (A \cdot \nabla)B \qquad [177]$$

(v) *The vector product of a vector and the curl of another vector.*

Here Eq. 32 is applied to the triple vector product

$$A \times \operatorname{curl} B = A \times [\nabla \times B] \qquad [178]$$

the vector function A being treated as a constant. Thus one obtains

$$A \times [\nabla \times B] = \nabla(A \cdot B)_A - (A \cdot \nabla)B \qquad [179]$$

or

$$A \times \operatorname{curl} B = \operatorname{grad}\,(A \cdot B)_A - (A \cdot \nabla)B \qquad [180]$$

(vi) *The gradient of a scalar product.*

This reads

$$\operatorname{grad}\,(A \cdot B) = \operatorname{grad}\,(A \cdot B)_A + \operatorname{grad}\,(A \cdot B)_B \qquad [181]$$

By Eq. 180,

$$\operatorname{grad}\,(A \cdot B)_A = A \times \operatorname{curl} B + (A \cdot \nabla)B \qquad [182]$$

$$\operatorname{grad}\,(A \cdot B)_B = B \times \operatorname{curl} A + (B \cdot \nabla)A \qquad [183]$$

Hence

$$\operatorname{grad}\,(A \cdot B) = A \times \operatorname{curl} B + B \times \operatorname{curl} A + (A \cdot \nabla)B + (B \cdot \nabla)A \qquad [184]$$

(vii) *The volume integral of the scalar product of a potential and a solenoidal vector function.*

A purely potential field is described by the vector function A, and a purely turbulent field by the vector B; that is,

$$\operatorname{curl} A = 0 \qquad [185]$$

and

$$\operatorname{div} B = 0 \qquad [186]$$

throughout all space occupied by these fields. In the volume integral

$$\int (A \cdot B)\, dv \qquad [187]$$

which is extended over all space, or over that portion to which the field B may be confined, the element of volume dv is represented by an elementary length ds of any one of the closed tubes characterizing the flow map for the solenoidal field B. If the cross-sectional area of a flow tube is denoted by da, the integral 187 may be written

$$\int (A \cdot B)\, dv = \int\int |B|\, da\, (A \cdot \mathbf{ds}) \qquad [188]$$

Here the vector element of length **ds** of a flow tube has the same direction as B. The integration with respect to ds extends around the closed circuit of a flow tube, and the integration with respect to da extends over all the flow tubes.

According to the definition of a flow tube, $|B|\,da$ is constant throughout the closed circuit mapped by this tube, so that this factor may be placed before the integral with respect to ds. Hence

$$\int (A \cdot B)\,dv = \int |B|\,da \oint A \cdot \mathbf{ds} \qquad [189]$$

Since A defines a potential field, it follows that

$$\oint A \cdot \mathbf{ds} = 0 \qquad [190]$$

so that the final result reads

$$\int (A \cdot B)\,dv = 0 \qquad [191]$$

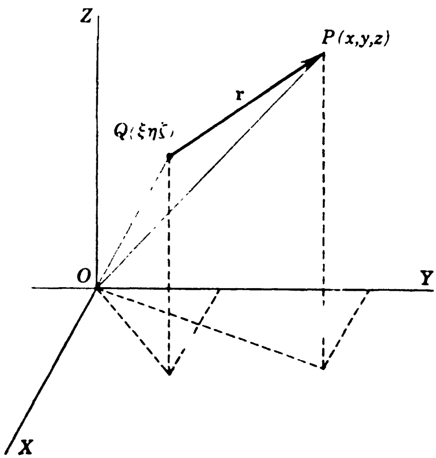


FIG. 17. The vector **r**.

17. The vector r

In connection with many field problems it is convenient to introduce a vector **r** which represents the vector distance between two points P and Q in space as shown in Fig. 17. The point Q may be the location of a cause (such as a source or vortex) whereas P is the point at which the resulting field is observed. The vector **r** is assumed to point from Q to P.

Hence if the co-ordinates of Q are denoted by ξ, η, ζ, and those of P by x, y, z, the components of $\mathbf{r}$ are given by

$$r_x = (x - \xi), \qquad r_y = (y - \eta), \qquad r_z = (z - \zeta) \tag{192}$$

and the vector $\mathbf{r}$ is

$$\mathbf{r} = i(x - \xi) + j(y - \eta) + k(z - \zeta) \tag{193}$$

The magnitude of $\mathbf{r}$ is expressed by the scalar function

$$r = \sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2} \tag{194}$$

In discussing various operations in terms of either the vector $\mathbf{r}$ or the scalar function r, a distinction must frequently be made according to which of the points P or Q is considered to be variable. If Q is fixed and P variable, the operations of differentiation or integration apply to the co-ordinates x, y, z. This state of affairs may be indicated by attaching a subscript P to the operator in question. Similarly, a subscript Q indicates that P is considered fixed, and the variables ξ, η, ζ, are affected by the operator. For example,

$$\text{grad}_P = i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z} \tag{195}$$

whereas

$$\text{grad}_Q = i\frac{\partial}{\partial \xi} + j\frac{\partial}{\partial \eta} + k\frac{\partial}{\partial \zeta} \tag{196}$$

Similar distinctions apply to the operator *div* and *curl.*

Because of the symmetrical manner in which the variables x, y, z on the one hand and ξ, η, ζ on the other, enter into the expressions 192 for the components of $\mathbf{r}$, the difference between a given operation as evaluated for the subscripts P and Q is readily recognized. Thus, for the operations, *grad*, *div*, *curl*, for example, the result for a subscript P is simply the negative of that for a subscript Q.

In the following more specific discussion, therefore, the variables are always assumed to be x, y, z. Without loss in the generality of this discussion, the fixed point Q may then be assumed to be coincident with the origin of co-ordinates. This simplifies the expressions 192, 193, and 194 to

$$r_x = x, \qquad r_y = y, \qquad r_z = z \tag{197}$$

$$\mathbf{r} = ix + jy + kz \tag{198}$$

and

$$r = \sqrt{x^2 + y^2 + z^2} \tag{199}$$

In terms of these expressions it is readily seen that

$$\operatorname{grad} r = i\frac{\partial r}{\partial x} + j\frac{\partial r}{\partial y} + k\frac{\partial r}{\partial z} = \frac{1}{r}(ix + jy + kz) = \frac{\mathbf{r}}{r} \qquad [200]$$

In other words, the gradient of the scalar function r is equal to a unit vector in the direction $\mathbf{r}$.

For the evaluation of the gradient of a function of r, it is useful to observe that

$$\operatorname{grad} = \frac{\partial}{\partial r}\left(i\frac{\partial r}{\partial x} + j\frac{\partial r}{\partial y} + k\frac{\partial r}{\partial z}\right) \qquad [201]$$

Hence if $f(r)$ denotes a function of r, and the partial derivative $\partial f/\partial r$ is written $f'(r)$,

$$\operatorname{grad} f(r) = f'(r) \operatorname{grad} r = f'(r)\frac{\mathbf{r}}{r} \qquad [202]$$

For example, if

$$f(r) = r^n \qquad [203]$$

then

$$\operatorname{grad} r^n = nr^{n-2}\mathbf{r} \qquad [204]$$

An application of this result to a commonly occurring form reads

$$\operatorname{grad}\left(\frac{1}{r}\right) = -\frac{\mathbf{r}}{r^3} \qquad [205]$$

According to Eq. 60, the divergence of the vector function $\mathbf{r}$ is

$$\operatorname{div} \mathbf{r} = \frac{\partial r_x}{\partial x} + \frac{\partial r_y}{\partial y} + \frac{\partial r_z}{\partial z} \qquad [206]$$

and with Eq. 197 this gives

$$\operatorname{div} \mathbf{r} = 3 \qquad [207]$$

From Eq. 165 it follows that

$$\operatorname{div}(r^n\mathbf{r}) = r^n \operatorname{div} \mathbf{r} + \mathbf{r} \cdot \operatorname{grad} r^n \qquad [208]$$

By substitution from Eqs. 204 and 207, this yields

$$\operatorname{div}(r^n\mathbf{r}) = (n + 3)r^n \qquad [209]$$

This result, together with Eq. 204, shows that

$$\operatorname{div} \operatorname{grad} r^n = \nabla^2 r^n = n(n + 1)r^{n-2} \qquad [210]$$

In particular, for $n = -1$, this gives the important result

$$\nabla^2\left(\frac{1}{r}\right) = 0 \qquad [211]$$

Evaluating the curl of the vector function $\mathbf{r}$ by substituting the components 197 into Eq. 104 shows that

$$\text{curl } \mathbf{r} = 0 \qquad [212]$$

and the use of Eq. 166 then shows further that

$$\text{curl } (r^n \mathbf{r}) = 0 \qquad [213]$$

In particular,

$$\text{curl } \left(\frac{\mathbf{r}}{r^2}\right) = 0 \qquad [214]$$

The vector function $\mathbf{r}$, or $(r^n\mathbf{r})$, may be looked upon as defining a pure potential field.

A number of vector relations formed from the function r in combination with an arbitrary vector function A are also useful. Using the result expressed by Eq. 184 together with Eq. 212 yields

$$\text{grad } (\mathbf{r} \cdot A) = \mathbf{r} \times \text{curl } A + (\mathbf{r} \cdot \nabla)A + (A \cdot \nabla)\mathbf{r} \qquad [215]$$

But

$$(A \cdot \nabla)\mathbf{r} = \left(A_x \frac{\partial}{\partial x} + A_y \frac{\partial}{\partial y} + A_z \frac{\partial}{\partial z}\right)\mathbf{r} = A \qquad [216]$$

so that

$$\text{grad } (\mathbf{r} \cdot A) = \mathbf{r} \times \text{curl } A + (\mathbf{r} \cdot \nabla)A + A \qquad [217]$$

Equations 173 and 212 show that

$$\text{div } [\mathbf{r} \times A] = -\mathbf{r} \cdot \text{curl } A \qquad [218]$$

Whereas Eqs. 177, 207, and 216 yield

$$\text{curl } [\mathbf{r} \times A] = \mathbf{r} \text{ div } A - 2A - (\mathbf{r} \cdot \nabla)A \qquad [219]$$

18. Curvilinear co-ordinates

Because of the geometry inherent in some physical problems, it is advantageous to designate various points in space in terms of co-ordinates other than the rectangular Cartesian ones which have been used in this discussion so far. In a problem which exhibits cylindrical symmetry, for example, it is usually effective to use cylindrical co-ordinates; if the physical problem has spherical symmetry, spherical polar co-ordinates are usually preferable, and so on.

The variables occurring in connection with any of these curvilinear co-ordinate systems may be denoted by u, v, w, as contrasted with the variables x, y, z, used for the rectangular Cartesian system. In the Cartesian system the curves for constant values of y and z, z and x, and

x and y form mutually orthogonal families (which in this instance are straight lines). For this reason, the rectangular Cartesian co-ordinates are said to be orthogonal.

In a curvilinear co-ordinate system, the curves for constant values of v and w, w and u, and u and v may also form mutually orthogonal families, for example, in the case of cylindrical or spherical polar co-ordinates. The curvilinear co-ordinates are then said to be orthogonal. The discussion of the present article is restricted to systems of this sort since the need for a more general system very rarely arises in practice. Moreover, the mathematical apparatus necessary for dealing with perfectly general curvilinear co-ordinates – the absolute differential calculus — requires a detailed study of considerable depth, which seems warranted only when sufficient use for it in connection with other practical problems presents itself.

For the orthogonal curvilinear co-ordinates, a set of mutually orthogonal unit vectors i_u, i_v, i_w are defined which are analogous to the vectors i, j, k in the Cartesian system. The unit vector i_u at any point in space is tangent to the curve v = constant, w = constant, at that point. Similarly, i_v is tangent to the curve w = constant, u = constant, and i_w is tangent to the curve u = constant, v = constant. The directions are, moreover, so chosen that i_u, i_v, i_w form a right-hand system.

It should be observed at this point that the scale of length which is implied in the designation of i_u, i_v, i_w as *unit* vectors is that pertaining to the rectangular Cartesian co-ordinate system. The scales for the curvilinear co-ordinate axes u, v, w at any point are in general different from this scale of length, and are moreover a function of the position of the point in space. In other words, the magnitudes of the increments du, dv, dw as measured in the Cartesian co-ordinate system are

$$\begin{aligned} ds_u &= e_u\,du \\ ds_v &= e_v\,dv \\ ds_w &= e_w\,dw \end{aligned} \qquad [220]$$

in which e_u, e_v, e_w are factors accounting for the differences in the scales for the u, v, and w co-ordinate axes at the point in question and the scale of length of the Cartesian system.

A given vector increment of length **ds** is expressed in the Cartesian co-ordinate system by

$$\mathbf{ds} = i\,dx + j\,dy + k\,dz \qquad [221]$$

and in the orthogonal curvilinear system by

$$\mathbf{ds} = i_u\,ds_u + i_v\,ds_v + i_w\,ds_w \qquad [222]$$

The length of the vector increment **ds** in terms of the Cartesian co-ordinates is determined by

$$(ds)^2 = (dx)^2 + (dy)^2 + (dz)^2 \tag{223}$$

whereas in terms of the curvilinear co-ordinates

$$(ds)^2 = (ds_u)^2 + (ds_v)^2 + (ds_w)^2 \tag{224}$$

or

$$(ds)^2 = e_u{}^2(du)^2 + e_v{}^2(dv)^2 + e_w{}^2(dw)^2 \tag{225}$$

it being understood that this length is measured by the scale of the Cartesian system.

The scale factors e_u, e_v, e_w, which in general are functions of the position of the point in question, that is, functions of x, y, z or u, v, w, may for given curvilinear co-ordinate system be found in several ways. In simple cases the expressions for ds_u, ds_v, and ds_w may be written down by inspection; whence the scale factors according to Eqs. 220 become evident. A more formal method of determining them follows.

For any particular curvilinear co-ordinate system, the co-ordinates x, y, z are expressible as functions of u, v, w, thus

$$x = f_1(u,v,w) \tag{226}$$

$$y = f_2(u,v,w) \tag{227}$$

$$z = f_3(u,v,w) \tag{228}$$

According to the rules of the differential calculus,

$$\begin{aligned}
\frac{\partial f_1}{\partial u}du + \frac{\partial f_1}{\partial v}dv + \frac{\partial f_1}{\partial w}dw &= dx \\
\frac{\partial f_2}{\partial u}du + \frac{\partial f_2}{\partial v}dv + \frac{\partial f_2}{\partial w}dw &= dy \\
\frac{\partial f_3}{\partial u}du + \frac{\partial f_3}{\partial v}dv + \frac{\partial f_3}{\partial w}dw &= dz
\end{aligned} \tag{229}$$

or

$$\begin{aligned}
\frac{\partial f_1}{\partial s_u}ds_u + \frac{\partial f_1}{\partial s_v}ds_v + \frac{\partial f_1}{\partial s_w}ds_w &= dx \\
\frac{\partial f_2}{\partial s_u}ds_u + \frac{\partial f_2}{\partial s_v}ds_v + \frac{\partial f_2}{\partial s_w}ds_w &= dy \\
\frac{\partial f_3}{\partial s_u}ds_u + \frac{\partial f_3}{\partial s_v}ds_v + \frac{\partial f_3}{\partial s_w}ds_w &= dz
\end{aligned} \tag{230}$$

In terms of the transformation matrices

$$\mathcal{J} = \begin{bmatrix} \dfrac{\partial f_1}{\partial u} & \dfrac{\partial f_1}{\partial v} & \dfrac{\partial f_1}{\partial w} \\[2ex] \dfrac{\partial f_2}{\partial u} & \dfrac{\partial f_2}{\partial v} & \dfrac{\partial f_2}{\partial w} \\[2ex] \dfrac{\partial f_3}{\partial u} & \dfrac{\partial f_3}{\partial v} & \dfrac{\partial f_3}{\partial w} \end{bmatrix} \tag{231}$$

and

$$\mathcal{J}_s = \begin{bmatrix} \dfrac{\partial f_1}{\partial s_u} & \dfrac{\partial f_1}{\partial s_v} & \dfrac{\partial f_1}{\partial s_w} \\[2ex] \dfrac{\partial f_2}{\partial s_u} & \dfrac{\partial f_2}{\partial s_v} & \dfrac{\partial f_2}{\partial s_w} \\[2ex] \dfrac{\partial f_3}{\partial s_u} & \dfrac{\partial f_3}{\partial s_v} & \dfrac{\partial f_3}{\partial s_w} \end{bmatrix} \tag{232}$$

the relations 229 and 230 may be written in the matrix form

$$\mathcal{J} \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \tag{233}$$

and

$$\mathcal{J}_s \begin{bmatrix} ds_u \\ ds_v \\ ds_w \end{bmatrix} = \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \tag{234}$$

From Eqs. 220, it is clear that

$$\begin{bmatrix} ds_u \\ ds_v \\ ds_w \end{bmatrix} = \begin{bmatrix} e_u & 0 & 0 \\ 0 & e_v & 0 \\ 0 & 0 & e_w \end{bmatrix} \times \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} \tag{235}$$

and hence that

$$\mathcal{J} = \mathcal{J}_s \begin{bmatrix} e_u & 0 & 0 \\ 0 & e_v & 0 \\ 0 & 0 & e_w \end{bmatrix} \tag{236}$$

Since both the Cartesian system and the curvilinear system are orthogonal, it follows (see Art. 4, Ch. III) that the matrix $\mathcal{J}_s$ in the transformation 234 is orthogonal, and hence that

$$(\mathcal{J}_s)_t = \mathcal{J}_s^{-1} \tag{237}$$

the subscript t denoting the transposed matrix. But from Eq. 236

$$\mathcal{F}_t = \begin{bmatrix} e_u & 0 & 0 \\ 0 & e_v & 0 \\ 0 & 0 & e_w \end{bmatrix} (\mathcal{F}_s)_t \qquad [238]$$

and with Eqs. 236 and 237 this gives

$$\mathcal{F}_t\mathcal{F} = \begin{bmatrix} e_u^2 & 0 & 0 \\ 0 & e_v^2 & 0 \\ 0 & 0 & e_w^2 \end{bmatrix} \qquad [239]$$

Hence, by means of Eq. 231 it follows that

$$e_u^2 = \left(\frac{\partial f_1}{\partial u}\right)^2 + \left(\frac{\partial f_2}{\partial u}\right)^2 + \left(\frac{\partial f_3}{\partial u}\right)^2 \qquad [240]$$

$$e_v^2 = \left(\frac{\partial f_1}{\partial v}\right)^2 + \left(\frac{\partial f_2}{\partial v}\right)^2 + \left(\frac{\partial f_3}{\partial v}\right)^2 \qquad [241]$$

$$e_w^2 = \left(\frac{\partial f_1}{\partial w}\right)^2 + \left(\frac{\partial f_2}{\partial w}\right)^2 + \left(\frac{\partial f_3}{\partial w}\right)^2 \qquad [242]$$

from which the scale factors are determined.

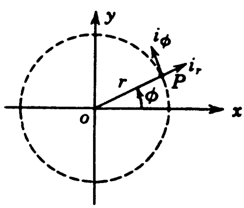


FIG. 18. Cylindrical co-ordinates with the z-axis as axis of cylindrical symmetry.

As an illustration, the cylindrical co-ordinates may be considered, for which the z-axis is chosen coincident with the axis of cylindrical symmetry (Fig. 18). The relations 226, 227, and 228 then read

$$x = r\cos\phi \qquad [243]$$

$$y = r\sin\phi \qquad [244]$$

$$z = z \qquad [245]$$

The variables r, ϕ, z are chosen to correspond respectively to u, v, w. Then Eqs. 240, 241, and 242 yield

$$e_r^2 = \cos^2\phi + \sin^2\phi + 0 = 1 \qquad [246]$$

$$e_\phi^2 = r^2\sin^2\phi + r^2\cos^2\phi + 0 = r^2 \qquad [247]$$

$$e_z^2 = 0 + 0 + 1 = 1 \qquad [248]$$

Hence

$$e_r = 1, \qquad e_\phi = r, \qquad e_z = 1 \qquad [249]$$

are the scale factors for this case. The vector increment of length is

$$\mathbf{ds} = i_r\,dr + i_\phi r\,d\phi + i_z\,dz \qquad [250]$$

that is,

$$ds_r = dr, \qquad ds_\phi = r\,d\phi, \qquad ds_z = dz \tag{251}$$

and

$$(ds)^2 = (dr)^2 + r^2(d\phi)^2 + (dz)^2 \tag{252}$$

In this simple example, the last three results can, of course, be written down at once from inspection of Fig. 18.

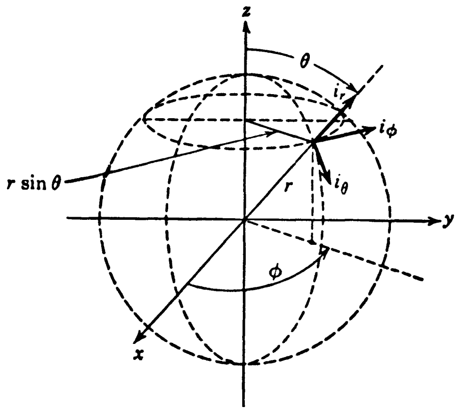


FIG. 19. Spherical polar coordinates.

The geometrical relations between spherical polar and rectangular co-ordinates are illustrated in Fig. 19. Accordingly, Eqs. 226, 227, and 228 become

$$x = r\sin\theta\cos\phi \tag{253}$$

$$y = r\sin\theta\sin\phi \tag{254}$$

$$z = r\cos\theta \tag{255}$$

The variables r, θ, ϕ are chosen to correspond respectively to u, v, w. Then Eqs. 240, 241, and 242 yield

$$e_r^2 = \sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi + \cos^2\theta = 1 \tag{256}$$

$$e_\theta^2 = r^2\cos^2\theta\cos^2\phi + r^2\cos^2\theta\sin^2\phi + r^2\sin^2\theta = r^2 \tag{257}$$

$$e_\phi^2 = r^2\sin^2\theta\sin^2\phi + r^2\sin^2\theta\cos^2\phi + 0 = r^2\sin^2\theta \tag{258}$$

Hence

$$e_r = 1, \qquad e_\theta = r, \qquad e_\phi = r\sin\theta \tag{259}$$

are the scale factors. The vector increment of length is*

$$\mathbf{ds} = i_r\,dr + i_\theta r\,d\theta + i_\phi r \sin\theta\,d\phi \qquad [260]$$

that is,

$$ds_r = dr, \qquad ds_\theta = r\,d\theta \qquad ds_\phi = r\sin\theta\,d\phi \qquad [261]$$

and

$$(ds)^2 = (dr)^2 + r^2(d\theta)^2 + r^2\sin^2\theta\,(d\phi)^2 \qquad [262]$$

The most commonly used vector operations are the gradient, divergence, curl, and div grad, or ∇^2. Formulation of these in terms of any orthogonal curvilinear co-ordinates u, v, w and specifically for the cylindrical and spherical polar co-ordinates are discussed in turn.

The expression for the gradient of a scalar function U reads

$$\text{grad } U = i_u \frac{\partial U}{\partial s_u} + i_v \frac{\partial U}{\partial s_v} + i_w \frac{\partial U}{\partial s_w} \qquad [263]$$

or

$$\text{grad } U = i_u \frac{1}{e_u}\frac{\partial U}{\partial u} + i_v \frac{1}{e_v}\frac{\partial U}{\partial v} + i_w \frac{1}{e_w}\frac{\partial U}{\partial w} \qquad [264]$$

With the help of Eqs. 249, this becomes, for *cylindrical co-ordinates,*

$$\text{grad } U = i_r \frac{\partial U}{\partial r} + i_\phi \frac{1}{r}\frac{\partial U}{\partial \phi} + i_z \frac{\partial U}{\partial z} \qquad [265]$$

and for *spherical polar co-ordinates,* substitution from Eqs. 259 gives

$$\text{grad } U = i_r \frac{\partial U}{\partial r} + i_\theta \frac{1}{r}\frac{\partial U}{\partial \theta} + i_\phi \frac{1}{r\sin\theta}\frac{\partial U}{\partial \phi} \qquad [266]$$

For the formulation of the general expression for the divergence of a vector function A, the definition given by Eq. 56 is applied to a curvilinear parallelepiped (Fig. 20) enclosed by the surfaces u = constant, and $u + du$ = constant; v = constant, and $v + dv$ = constant; w = constant, and $w + dw$ = constant. Except for differentials of higher order, the elementary areas of opposite faces of this parallelepiped are equal. Specifically, the elementary areas of the faces normal to i_u, i_v, i_w are given respectively by

$$da_u = ds_v\,ds_w = e_v e_w\,dv\,dw \qquad [267]$$

$$da_v = ds_w\,ds_u = e_w e_u\,dw\,du \qquad [268]$$

$$da_w = ds_u\,ds_v = e_u e_v\,du\,dv \qquad [269]$$

*In this example, one may also write down the expression for **ds** directly and thus obtain the scale factors without applying the formal method.

The volume of the parallelepiped is

$$dV = ds_u\, ds_v\, ds_w = e_u e_v e_w\, du\, dv\, dw \qquad [270]$$

The surface integral of A is given by

$$\oint A \cdot \mathbf{da} = [A_u da_u]_{u+du} - [A_u da_u]_u + [A_v da_v]_{v+dv} - [A_v da_v]_v + [A_w da_w]_{w+dw} - [A_w da_w]_w \qquad [271]$$

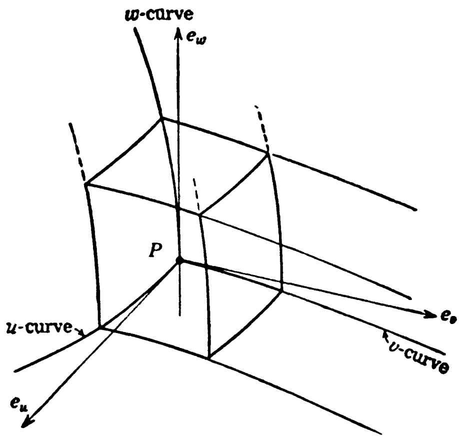


FIG. 20. A curvilinear parallelepiped formed by surfaces of constant u, v, and w.

in which A_u, A_v, A_w are the components of A in the directions of i_u, i_v, i_w respectively, and the subscripts indicate that the quantity enclosed by the brackets is to be evaluated at the points $u + du, v, w$; u, v, w; $u, v + dv, w$; u, v, w; etc. Now

$$[A_u\, da_u]_{u+du} = [A_u\, da_u]_u + \frac{\partial}{\partial u}[A_u\, da_u]_u\, du \qquad [272]$$

so that

$$[A_u\, da_u]_{u+du} - [A_u\, da_u]_u = \frac{\partial}{\partial u}[A_u\, da_u]_u\, du$$

$$= \frac{\partial}{\partial u}[A_u e_v e_w]_u\, du\, dv\, dw \qquad [273]$$

and so forth. Hence

$$\oint A \cdot \mathbf{da} = \left\{\frac{\partial}{\partial u}(A_u e_v e_w) + \frac{\partial}{\partial v}(A_v e_w e_u) + \frac{\partial}{\partial w}(A_w e_u e_v)\right\} du\, dv\, dw \qquad [274]$$

so that the general expression for the divergence becomes

$$\operatorname{div} A = \frac{\oint A \cdot \mathbf{da}}{ds_u\, ds_v\, ds_w}$$

$$= \frac{1}{e_u e_v e_w}\left\{\frac{\partial}{\partial u}(A_u e_v e_w) + \frac{\partial}{\partial v}(A_v e_w e_u) + \frac{\partial}{\partial w}(A_w e_u e_v)\right\} \qquad [275]$$

In *cylindrical co-ordinates*, this reads

$$\operatorname{div} A = \frac{1}{r}\left\{\frac{\partial}{\partial r}(A_r r) + \frac{\partial}{\partial \phi}(A_\phi) + \frac{\partial}{\partial z}(A_z r)\right\} \qquad [276]$$

or

$$\operatorname{div} A = \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{1}{r}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z} \qquad [277]$$

In *spherical polar co-ordinates*,

$$\operatorname{div} A = \frac{1}{r^2 \sin\theta}\left\{\frac{\partial}{\partial r}(A_r r^2 \sin\theta) + \frac{\partial}{\partial \theta}(A_\theta r \sin\theta) + \frac{\partial}{\partial \phi}(A_\phi r)\right\} \qquad [278]$$

or

$$\operatorname{div} A = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 A_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(A_\theta \sin\theta) + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi} \qquad [279]$$

The expressions for the components of the curl of the vector function A are obtained through applying the defining relation 89 to elementary curvilinear rectangles lying in planes normal to the directions of i_u, i_v, and i_w. Except for differentials of higher order, the opposite sides of such a rectangle are equal. Considering the u-component, the sides of the rectangle are $ds_v = e_v\, dv$ and $ds_w = e_w\, dw$. The line integral of A taken around this rectangle is expressed by

$$\oint_u A \cdot \mathbf{ds} = [A_v\, ds_v]_w + [A_w\, ds_w]_{v+dv} - [A_v\, ds_v]_{w+dw} - [A_w\, ds_w]_v \qquad [280]$$

But

$$[A_w\, ds_w]_{v+dv} = [A_w\, ds_w]_v + \frac{\partial}{\partial v}[A_w\, ds_w]_v\, dv \qquad [281]$$

and

$$[A_v\, ds_v]_{w+dw} = [A_v\, ds_v]_w + \frac{\partial}{\partial w}[A_v\, ds_v]_w\, dw \qquad [282]$$

so that Eq. 280 becomes

$$\oint_u A \cdot \mathbf{ds} = \frac{\partial}{\partial v}[A_w\, ds_w]_v\, dv - \frac{\partial}{\partial w}[A_v\, ds_v]_w\, dw \qquad [283]$$

or

$$\oint_u A \cdot \mathbf{ds} = \left\{\frac{\partial}{\partial v}(A_w e_w)_v - \frac{\partial}{\partial w}(A_v e_v)_w\right\} dv\, dw \qquad [284]$$

The enclosed area is $da_u = e_v e_w\, dv\, dw$. Hence by Eq. 89,

$$\mathrm{curl}_u\, A = \frac{1}{e_v e_w}\left\{\frac{\partial}{\partial v}(A_w e_w) - \frac{\partial}{\partial w}(A_v e_v)\right\} \qquad [285]$$

and similarly

$$\mathrm{curl}_v\, A = \frac{1}{e_w e_u}\left\{\frac{\partial}{\partial w}(A_u e_u) - \frac{\partial}{\partial u}(A_w e_w)\right\} \qquad [286]$$

$$\mathrm{curl}_w\, A = \frac{1}{e_u e_v}\left\{\frac{\partial}{\partial u}(A_v e_v) - \frac{\partial}{\partial v}(A_u e_u)\right\} \qquad [287]$$

In *cylindrical co-ordinates*, these expressions become

$$\mathrm{curl}_r\, A = \frac{1}{r}\frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z} \qquad [288]$$

$$\mathrm{curl}_\phi\, A = \frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r} \qquad [289]$$

$$\mathrm{curl}_z\, A = \frac{1}{r}\frac{\partial}{\partial r}(rA_\phi) - \frac{1}{r}\frac{\partial A_r}{\partial \phi} \qquad [290]$$

and in *spherical polar co-ordinates*, they read

$$\mathrm{curl}_r\, A = \frac{1}{r \sin\theta}\frac{\partial}{\partial \theta}(A_\phi \sin\theta) - \frac{1}{r\sin\theta}\frac{\partial A_\theta}{\partial \phi} \qquad [291]$$

$$\mathrm{curl}_\theta\, A = \frac{1}{r\sin\theta}\frac{\partial A_r}{\partial \varphi} - \frac{1}{r}\frac{\partial}{\partial r}(rA_\phi) \qquad [292]$$

$$\mathrm{curl}_\phi\, A = \frac{1}{r}\frac{\partial}{\partial r}(rA_\theta) - \frac{1}{r}\frac{\partial A_r}{\partial \theta} \qquad [293]$$

The evaluation of $\nabla^2 U$ is obtained through substituting the components of grad U from Eq. 264 for the components of A in Eq. 275. This gives

$$\nabla^2 U = \frac{1}{e_u e_v e_w}\left\{\frac{\partial}{\partial u}\left(\frac{e_v e_w}{e_u}\frac{\partial U}{\partial u}\right) + \frac{\partial}{\partial v}\left(\frac{e_w e_u}{e_v}\frac{\partial U}{\partial v}\right) + \frac{\partial}{\partial w}\left(\frac{e_u e_v}{e_w}\frac{\partial U}{\partial w}\right)\right\} \qquad [294]$$

For *cylindrical co-ordinates*, this yields

$$\nabla^2 U = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial U}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 U}{\partial \phi^2} + \frac{\partial^2 U}{\partial z^2} \qquad [295]$$

In *spherical polar co-ordinates*,

$$\nabla^2 U = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial U}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 U}{\partial \phi^2} \qquad [296]$$

The first term on the right-hand side of this equation may be written in other forms by noting that

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = \frac{\partial^2 U}{\partial r^2} + \frac{2}{r}\frac{\partial U}{\partial r} = \frac{1}{r}\frac{\partial^2 (rU)}{\partial r^2} \qquad [297]$$

PROBLEMS

1. Let A, B, C be three vectors extending from a fixed point, O, to the points a, b, c, respectively. Express the directed segments $\overrightarrow{ab}$ and $\overrightarrow{ac}$ in terms of A, B, C. If b lies on the line ac and divides it in the ratio

$$\frac{ab}{ac} = m$$

write the expression for B in terms of A, C, and m. Conversely, if $B = mA + nC$, where m and n are scalars such that $m + n = 1$, show that b must lie on ac and find the ratio

$$\frac{ab}{ac}$$

2. Let A, B, C, D be four vectors extending from a fixed point O to the points a, b, c, d, respectively. As in Prob. 1, show that the necessary and sufficient conditions yielding a, b, c, d coplanar are expressed by the relations

$$\alpha + \beta + \gamma = 1$$

$$D = \alpha A + \beta B + \gamma C$$

in which α, β, γ are scalars.

3. The vectors from the origin to the points a, b, c, d are

$$A = i + j + k$$

$$B = i2 + j3$$

$$C = i3 + j5 - k2$$

$$D = -j + k$$

Express the vectors $\overrightarrow{ab}, \overrightarrow{bc}, \overrightarrow{cd}, \overrightarrow{da}$ in terms of the unit vectors i, j, k. Show that $\overrightarrow{ab}$ and $\overrightarrow{cd}$ are parallel.

4. Prove that the absolute value or length of a vector A is given by

$$A = \sqrt{A \cdot A}$$

Find the lengths of the vectors A, B, C, and D in Prob. 3.

5. Evaluate the following scalar products:

$$\text{(i)}\quad (i3 + j4 + k2) \cdot (j - k5)$$
$$\text{(ii)}\quad (i + j2 - k3) \cdot (i4 + j + k2)$$
$$\text{(iii)}\quad (i50 + k35) \cdot (-i3 - j17 + k2)$$

6. Verify Eq. 3, page 188, for the case where

$$A = i2 + j3 - k$$
$$B = -i - j + k3$$
$$C = i - j2$$
$$D = -i2 + j + k3$$

7. Find the cosine of the angle between the vectors

$$A = i - j2 - k2$$
$$B = i2 + j - k2$$

8. The sides of a triangle are vectors A, B, and C, of length a, b, and c, respectively, such that

$$A = B - C$$

By using scalar multiplication, square both sides of this equation and, by interpreting the result geometrically, obtain the law of cosines.

9. If A and B are nonzero vectors, show that the necessary and sufficient condition for A and B to be:

(i) parallel is $A \times B = 0$

(ii) perpendicular is $A \cdot B = 0$

10. Evaluate the following vector products:

$$\text{(i)}\quad (i2 + j2 - k) \times (i + j)$$
$$\text{(ii)}\quad (i5 - j + k) \times (i2 - j + k2)$$
$$\text{(iii)}\quad (i + j + k) \times (i + j + k2)$$

11. If A and B are sides of a triangle, show that the area of the triangle is given by

$$\text{area} = \tfrac{1}{2}\sqrt{(A \times B) \cdot (A \times B)}$$

Calculate the area when

$$A = i + j + k$$
$$B = i2 + k3$$

12. Prove Lagrange's identity

$$(A \times B) \cdot (C \times D) = \begin{vmatrix} (A \cdot C) & (A \cdot D) \\ (B \cdot C) & (B \cdot D) \end{vmatrix}$$

Use this to find an alternative form for the area formula given in Prob. 11.

13. Using the vectors given in Prob. 10 (iii), compute V_x, V_y, V_z, and the projections $A^{(yz)}$, $A^{(zx)}$, $A^{(xy)}$ of Eq. 14, page 191. Find the angles θ_{yz}, θ_{zx}, θ_{xy}, and in terms of these results verify Eqs. 16, 17, and 18, page 191.

14. If A, B, C are coterminous edges of a tetrahedron, show that the volume of the tetrahedron is given by

$$\text{volume} = \tfrac{1}{6}|A \times B \times C|$$

Calculate the volume when

$$A = i3 - j + k$$
$$B = i + j + k$$
$$C = i - k$$

15. Show that

$$\begin{aligned}(A \times B) \times (C \times D) &= (A \times B \cdot D)C - (A \times B \cdot C)D \\ &= (A \times C \cdot D)B - (B \times C \cdot D)A\end{aligned}$$

16. Prove the formula

$$(A \times B) \cdot (B \times C) \times (C \times A) = (A \times B \cdot C)^2$$

17. If $U = x^2 + 4y^2 + 16z^2$, find grad U at the point $P(2,2,1)$. What is the shape of the constant value surface through P? Find the directional derivative of U,

$$\frac{\partial U}{\partial s}$$

at P, if the direction of s at P is along the vector

$$A = -i + j2 + k2$$

18. Verify Eq. 48, page 198, if

$$U = x^2 + yz$$
$$V = xyz$$

19. Find a function $U(x,y,z)$ with gradient equal to

$$\text{grad } U = i2x + j5y^4 + k4z^3$$

20. In fluid dynamics, there arises the energy function

$$\phi(x,y,z) = \int_{p_0}^{p} \frac{dp}{\rho}$$

where the density of the fluid $\rho(x,y,z)$ is a function of the pressure $p(x,y,z)$. Find the general expression for

$$\text{grad } \phi$$

If in particular

$$p = (x^2 + y^2 + z^2) - \tfrac{1}{2}$$

and

$$\rho = cp \qquad (c, \text{a constant})$$

show that

$$\text{grad}\,\phi = -\frac{1}{c}\frac{ix + jy + kz}{(x^2 + y^2 + z^2)}$$

21. Let $U = x^2 + 4y^2 + 16z^2$ define a potential. Calculate the line integral

$$\int \text{grad}\,U \cdot \mathbf{ds}$$

from $P_1(1,1,1)$ to $P_2(2,2,2)$:

(i) By integrating along straight lines from $(1,1,1)$ to $(2,1,1)$ to $(2,2,1)$ to $(2,2,2)$.
(ii) By integrating along the straight line P_1P_2.

22. Calculate the divergence of the following vector fields at the points indicated. Which of the points represent sources? Sinks?

(i) $A = i(x^2 - y^2) + j(z^2 - x^2) + k(y^2 - z^2)$ at $(2,3,1)$
(ii) $A = (ix + jy + kz)(x^2 + y^2 + z^2)$ at $(-1,2,2)$
(iii) $A = (i3z - j8x + k5y)(xy)$ at $(1,0,0)$

23. By considering the defining equation 56, page 201, calculate div $(ix + jy + kz)$ at the origin, using a sphere of radius Δr, with center at the origin, as the infinitesimal volume.

24. Determine the volume of fluid per second flowing out of a spherical region of radius 3 feet if the vector velocity field for the region is given by

$$V = ix^3 + jy^3 + kz^3 \quad \text{feet per second}$$

referred to a set of axes with origin at the center of the sphere.

25. Calculate the volume of fluid per second flowing into a cube measuring two feet along each edge if the vector velocity field is given by

$$V = -ix^3y^2 + jy^2z + kz^2x \quad \text{feet per second}$$

referred to a set of axes with origin at the center of the cube and such that the x-, y- and z-axes are perpendicular to a face of the cube.

26. Show that the volume enclosed by a closed surface S is

$$\text{volume} = \tfrac{1}{6}\int_S \nabla(r^2)\cdot \mathbf{da}$$

where r is the distance from the origin to the variable point; that is, $r = \sqrt{x^2 + y^2 + z^2}$.

27. Find the potential $U(x,y,z)$ associated with a vector force field which is directed toward the origin, with magnitude inversely proportional to the distance from the origin. (*Hint:* By symmetry, U will be a function only of r, the distance from the origin.)

28. The velocity of a fluid is radially outward from a point source and is proportional to the distance from the source. Find the velocity potential associated with this vector velocity field.

29. A sphere of radius a with center at the origin contains a space charge of density

$$\rho = a^2 - r^2 \qquad (\text{where } r = \sqrt{x^2 + y^2 + z^2})$$

There are no charges exterior to the sphere so that the resulting field will possess

spherical symmetry, that is,

$$A = f(r)\mathbf{r} \qquad (\text{where } \mathbf{r} = ix + jy + kz)$$

Using Eq. 64, page 205, on a sphere of radius r, evaluate $f(r)$ and thus show that

$$A = \frac{8\pi}{15}\frac{a^5}{r^3}\mathbf{r} \qquad (\text{when } r > a)$$

and

$$A = 4\pi\frac{a^2}{3} - \frac{r^2}{5}\mathbf{r} \qquad (\text{when } r < a)$$

30. In the preceding problem the potential U associated with the field A may be computed from

$$U = \int_r^\infty A \cdot d\mathbf{r}$$

(This may be thought of as the work done by the field when a unit charge, upon which it acts, moves from r out to infinity.) Calculate the expression for U valid outside the sphere and show that U satisfies

$$\nabla^2 U = 0$$

Also find the potential U valid for points inside the sphere and show that here U satisfies

$$\nabla^2 U = -4\pi\rho$$

31. On the surface of a sphere of radius a and center at the origin is a surface charge of constant density σ. By the same procedure used in Probs. 29 and 30, determine the vector field A, and the scalar potential U, for points interior and exterior to the sphere. Show that Laplace's equation is satisfied for all points except on the surface of the sphere, and that on the surface itself, Eq. 68, page 206, is satisfied.

32. Calculate the curl of each of the following vector fields:

$$\text{(i)}\ A = -iy + jx$$

$$\text{(ii)}\ A = i3xy + j2yx + kyz$$

$$\text{(iii)}\ A = ixyz + j(x^2 + y^2 + z^2) + k(x + y + z)$$

33. Verify Stokes's theorem by computing separately

$$\oint_L A \cdot d\mathbf{s}$$

and

$$\int_S (\text{curl } A) \cdot d\mathbf{a}$$

in the case where

$$A = iy - jx$$

and the surface S is the hemisphere of radius b, above the xy-plane, whose boundary

L is the circle

$$x^2 + y^2 = b^2$$

in the xy-plane.

34. Evaluate

$$\oint A \cdot d\mathbf{s}$$

around a square of side b, which has a corner at the origin, one side on the x-axis, and one side on the y-axis, if

$$A = i(x^2 - y^2) + j2xy$$

35. Consider an infinitely long, straight vortex thread of constant moment I. By using Eqs. 102 and 103, page 213, together with the fact that the resulting field will possess cylindrical symmetry (that is, the lines of force will be circles about the vortex), deduce the expression for the vector field A. Use cylindrical co-ordinates with the z-axis along the vortex.

Show that the potential U which satisfies Eq. 130, page 220, is given by

$$U = 2I(\theta_0 - \theta)$$

where θ is the angle in cylindrical co-ordinates. Verify that this is a multivalued potential function which satisfies Eq. 136, page 222.

36. Consider an infinitely long right circular cylinder of radius b whose axis is the z-axis of cylindrical co-ordinates. Interior to the cylinder is a vortex density J defined as follows; the direction of J is always parallel to the z-axis and the magnitude is given by

$$|J| = r^2$$

The exterior of the cylinder is assumed to be vortex-free. Using the same sort of reasoning as in the preceding problem, determine the vector field A, obtaining expressions valid inside and outside the cylinder. For the inside of the cylinder, find the vector potential P which satisfies

$$\text{curl } P = A$$

$$\text{div } P = 0$$

(*Hint:* From Eqs. 126 and 127, page 219, it can be seen that the direction of P is the same as that of J.)

For the outside of the cylinder, find the scalar potential U which satisfies

$$\text{grad } U = -A$$

37. A point $P(x,y,z)$ moves along a space curve in such a way that its co-ordinates are given by the following functions of time:

$$x = a \cos t$$

$$y = a \sin t$$

$$z = bt$$

Find the vector velocity and acceleration of the point P.

38. The potential energy of the region in the preceding problem is given by

$$U = \frac{z}{x^2 + y^2} \sin t$$

Find dU/dt for the point $P(x,y,z)$ of Prob. 37, at time t.

39. With reference to Probs. 37 and 38, let $A = -\text{grad } U$. Determine dA/dt for the same $P(x,y,z)$ at time t.

40. Let

$$U = \frac{x^2 + y^2}{z}$$

$$A = i(x^2 + yz) + j(y^2 + zx) + k(z^2 + xy)$$

$$B = i(xy) + j(yz) + k(zx)$$

Calculate:

(i) grad U
(ii) div A, div B
(iii) curl A, curl B
(iv) div (UA)
(v) curl (UA)
(vi) div $(A \cdot B)$
(vii) curl $(A \times B)$
(viii) grad $(A \cdot B)$

41. If A is a constant vector, prove

$$\nabla(A \cdot \mathbf{r}) = A$$

42. Let V_1 and V_2 be vectors from the fixed points (x_1,y_1,z_1), (x_2,y_2,z_2) to the variable point (x,y,z). Show that

(i) div $(V_1 \times V_2) = 0$
(ii) curl $(V_1 \times V_2) = 2(V_1 - V_2)$
(iii) grad $(V_1 \cdot V_2) = V_1 + V_2$

43. If A is an arbitrary vector field, evaluate

$$(A \times \nabla) \cdot R$$

and

$$(A \times \nabla) \times R$$

44. Obtain Eq. 287, page 243, directly from

$$\nabla^2 U(x,y,z) = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2}$$

by the substitution given by Eqs. 235 to 237, page 237.

45. Obtain Eq. 288, page 243, directly as in Prob. 44 by means of the substitution given by Eqs. 245 to 247, page 238.

46. Suppose that

$$\begin{aligned}\psi &= \psi(\alpha_1)\\ \alpha_1 &= \psi_1(\alpha_2)\\ \alpha_2 &= \psi_2(\alpha_3)\\ &\cdots\cdots\cdots\cdots\\ \alpha_n &= \alpha_n(u, v, w)\end{aligned}$$

a series of relations through which ψ is expressed as a function of u, v, w. Show that

$$\operatorname{grad} \psi = \nabla\psi = \frac{\partial\psi}{\partial\alpha_1}\frac{\partial\alpha_1}{\partial\alpha_2}\frac{\partial\alpha_2}{\partial\alpha_3}\cdots\frac{\partial\alpha_{n-1}}{\partial\alpha_n}\nabla\alpha_n$$

47. Apply the formula of the preceding problem to compute the gradient of

$$\psi = \ln \sin e^{\sinh (A\cdot B)}$$

in which

$$A = i_u a_1 + i_v a_2 + i_w a_3$$

$$B = i_u u + i_v v + i_u w$$

the quantities a_1, a_2, a_3 being constants.

48. Consider a rectangular co-ordinate system with origin at the point O and let Q and P be two other points separated from O by the distances ρ and R respectively. Let $\mathbf{r}$ denote the vector distance from Q to P (as defined in Art. 17) and similarly let $\boldsymbol{\rho}$ and $\mathbf{R}$ denote the vectors OQ and OP. The angle between the latter two vectors is ψ. If Q is assumed fixed while P moves upon the surface of a sphere with radius R, compute the magnitude and the direction cosines of the vector $\operatorname{grad}_P \cos\psi$. Show that this vector is perpendicular to $\mathbf{R}$.

49. With reference to the previous problem compute the magnitude and direction cosines of the vector $\operatorname{grad}_Q \cos\psi$ when P is fixed and Q moves over the surface of a sphere of radius ρ. Which vector, $\boldsymbol{\rho}$ or $\mathbf{R}$, is perpendicular to $\operatorname{grad}_Q \cos\psi$?

50. If $\phi(r)$ is some scalar function of the distance r as defined in Prob. 48, show that

$$\operatorname{grad}_P \phi = \frac{\partial\phi}{\partial r}\operatorname{grad}_P r = -\operatorname{grad}_Q r$$

Compute the magnitude and the direction cosines of this vector if

$$\phi = \frac{e^{-jkr}}{r} \qquad (k \text{ being a constant})$$

51. Referring to the statement of the previous problem, show that

$$\nabla_Q^2\phi = \nabla_P^2\phi = -k^2\phi$$

52. If $\mathbf{a}$ is a constant vector and $\phi(r)$ a scalar function of r, show that

$$\nabla_Q \times (\mathbf{a}\phi) = \mathbf{a} \times \nabla_P\phi$$

and

$$\nabla_Q \times \nabla_Q \times (\mathbf{a}\phi) = \nabla_Q(\mathbf{a}\cdot\nabla_Q\phi) - \mathbf{a}\nabla_Q\phi$$

53. Prove the vector identity

$$E \equiv \frac{1}{A\cdot A}[A \times E \times A + (E\cdot A)A]$$

in which E and A are any vectors.

54. Prove the following vector relations:

$$E \cdot (\nabla \times \nabla \times \mathbf{a}\phi) = k^2\phi E \cdot \mathbf{a} + E \cdot \nabla(\mathbf{a} \cdot \nabla\phi)$$

$$E \cdot \nabla(\mathbf{a} \cdot \nabla\phi) = \nabla \cdot [(\mathbf{a} \cdot \nabla\phi)E] - (\mathbf{a} \cdot \nabla\phi)(\nabla \cdot E)$$

$$E \cdot (\nabla \times \nabla \times \mathbf{a}\phi) = \mathbf{a} \cdot \{k^2\phi E - (\nabla \cdot E)\nabla\phi\} + \nabla \cdot [(\mathbf{a} \cdot \nabla\phi)E]$$

in which E is a vector function of x, y, z, ϕ is the scalar function defined in Prob. 50, and $\mathbf{a}$ is a constant vector.

55. If A and B are any two vector functions of the space co-ordinates, show by means of Gauss's law that

$$\int_V (B \cdot \nabla \times \nabla \times A - A \cdot \nabla \times \nabla \times B)\, dv = \int_S [A \times \nabla \times B - B \times \nabla \times A] \cdot \mathbf{da}$$

in which **da** is a vector element of the surface S and dv is an element of the enclosed volume V.

CHAPTER VI

Functions of a Complex Variable

It is assumed that the reader has some acquaintance with the subject of complex numbers and the representation of them in the complex plane (also known as the *Gaussian plane*). Here it is customary to consider the x-axis as the " axis of reals " and the y-axis as the " axis of imaginaries." The symbol $j = \sqrt{-1}$ prefixed to a real number signifies that the latter is the imaginary or y-component of a given complex number. The y-axis is, therefore, sometimes referred to as the "j-axis."

A complex number $z = x + jy$ is plotted in the complex plane as the point $P(x,y)$, and z is interpreted geometrically as a vector* drawn from the origin to this point. This interpretive procedure is referred to as the *rectangular* representation of the complex number (or vector) z. Its polar representation is given by $z = \rho e^{j\theta}$, whence $\rho = \sqrt{x^2 + y^2}$, and $\theta = \tan^{-1} (y/x)$. The familiar law of parallelogram addition applies to the addition of any given set of complex numbers. The details of this process as well as those involved in the multiplication and division of complex numbers are not further elaborated upon here.

The object of the discussion in these pages is rather to give the reader some acquaintance with a complex function whose independent variable is a complex number. The most obvious novelty exhibited by the function of a complex variable, contrasted with a real function of a real variable, lies in the fact that the values of both the function and the variable are no longer characterized by single numbers; two numbers are now required for the specification of each. Thus a value of the variable z involves the specification of the two quantities x and y, and since the function of z is likewise complex, its values also involve the specification of a real and an imaginary part.

In view of this situation it is clear that the process of graphical representation requires new methods in the case of complex functions. In addition, one should carefully review the fundamental operations of the differential and integral calculus in order to see whether their familiar interpretation may be extended in some way to apply to complex functions. This extension should be made with the minimum possible change in basic conceptions in order that the many useful relations known with

*Complex numbers are frequently referred to by electrical engineers as vectors. It has been pointed out that the term " vectors " is in this connection misused and that such misuse may lead one to draw false conclusions or otherwise fall into dangerous byways of thought. This view is not shared in this book, principally because the discussion given in Art. 5 shows that it has little justification.

regard to functions of a real variable may become available also to the manipulation of the complex ones.*

The most useful operations familiar to the reader in connection with functions of a real variable are those of differentiation, integration, and expansion in series. A discussion of the extension of these conceptions to functions of a complex variable is the principal aim of the following articles.

In these discussions no particular attempt is made to give rigorous proofs. The derivations are given entirely for the purpose of providing the reader with a partial insight into the relevant fundamental interrelationships. Wherever possible, an attempt is made to establish contact between the new conceptions and those with which the student has an acquaintance of longer standing. To the engineer who is trying to gain a working knowledge of function theory, rigorous proofs are a waste of effort, but plausibility arguments do serve a useful purpose in that they provide the necessary circumspection for facile and intelligent use of this important mathematical tool, and at the same time lay the groundwork for a more thorough study which may seem desirable at a later time.

1. Differentiation

The complex function of a complex variable z is denoted by

$$w = f(z) \tag{1}$$

with

$$z = x + jy \tag{2}$$

and

$$w = u + jv \tag{3}$$

Therefore u and v are functions of both x and y; thus

$$u = u(x,y), \qquad v = v(x,y) \tag{4}$$

It should be observed that both the real and imaginary parts u and v are real functions of the two real independent variables x and y, which are the real and imaginary parts of the complex variable z. The language used here may be somewhat confusing to the reader, inasmuch as v and y

*This approach to the consideration of the theory of functions of a complex variable may appear to some readers to be somewhat strange. It should be remembered, however, that these pages are addressed primarily to the engineering student whose previous experience with mathematics has been confined almost wholly to real functions of a real variable. To him the process of regarding the present discussions as an extension of some of the manipulations which apply to functions of a real variable not only appears to be sensible but also is the course which his process of learning will take in any case.

are spoken of as the " imaginary " parts of w and z respectively, whereas at the same time they are pointed out as being real quantities. The strangeness of this method of expression should, however, readily be overcome by concentration upon its mathematical significance as expressed by the relations 1 to 4.

It is now of interest to examine whether the derivative of the function w with respect to the variable z may be interpreted as the limit of the ratio $\Delta w/\Delta z$, in which Δw and Δz represent corresponding increments, as the increment Δz approaches zero. Reflecting upon this situation at once discloses an apparent difficulty, since one is reminded of the fact that

$$\Delta z = \Delta x + j\,\Delta y \tag{5}$$

and hence that Δz may be interpreted in an infinity of ways. If one assumes for the moment that the increment Δz has a fixed magnitude, its direction in the complex plane may be varied in an infinite number of ways, thus yielding an infinite number of corresponding increments Δw in the function. It does not necessarily follow, of course, that the ratio $\Delta w/\Delta z$ correspondingly assumes an infinite number of values, but unless, in the limit $\Delta z \to 0$, this ratio is independent of the direction of Δz, the derivative of the function w evidently does not possess a unique value.

Whereas it is conceivable that the extension of the usual conception of a derivative to functions of a complex variable may require a distinction with regard to the direction assigned to the increment Δz in the complex plane, the simplicity and general usefulness of this derivative would unquestionably be greatly impaired if its value were subjected to such a restriction. The undesirability of the latter suggests that one ought to demand that the derivative operation be completely free from this restriction, and then inquire into the bonds which are thereby laid upon the nature of the complex function. If these are not so severe as to rule out of consideration such classes or kinds of functions as one would like to see embraced by the theory which is the object of this discussion, one may still be served by pursuing it.

As will be seen shortly, it turns out that this point of view may well be taken and, surprisingly enough, that more functions out of a set written down at random fall into a classification which meets these bonds than one might at the outset expect. In fact, the results are so gratifying in this respect that one is justified in ruling out of further consideration all complex functions which do not comply with this demand and in stipulating that the term " function of a complex variable " shall apply only to those that do.

The conditions under which the derivative of the function w with respect to the complex variable z may be independent of the direction in which the increment Δz is taken are readily found by first indicating the

derivative by means of the relations 2 and 3 as

$$\frac{dw}{dz} = \frac{du + j\,dv}{dx + j\,dy} \tag{6}$$

Now from Eq. 4

$$du = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy \tag{7}$$

and

$$dv = \frac{\partial v}{\partial x}dx + \frac{\partial v}{\partial y}dy \tag{8}$$

Substituting into Eq. 6 and writing the result in the form

$$\frac{dw}{dz} = \frac{\left(\dfrac{\partial u}{\partial x} + j\dfrac{\partial v}{\partial x}\right) + \left(\dfrac{\partial u}{\partial y} + j\dfrac{\partial v}{\partial y}\right)\dfrac{dy}{dx}}{1 + j\dfrac{dy}{dx}} \tag{9}$$

show that the direction of the differential increment dz is determined by dy/dx, and hence, if the expression 9 is to be independent of this direction, that the necessary conditions are expressed by

$$\frac{\dfrac{\partial u}{\partial x} + j\dfrac{\partial v}{\partial x}}{\dfrac{\partial u}{\partial y} + j\dfrac{\partial v}{\partial y}} = \frac{1}{j} \tag{10}$$

or

$$\frac{\partial u}{\partial y} + j\frac{\partial v}{\partial y} = -\frac{\partial v}{\partial x} + j\frac{\partial u}{\partial x} \tag{11}$$

Equating real and imaginary parts in this equation gives the conditions

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \tag{12}$$

and

$$\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \tag{13}$$

These are the conditions which the real and imaginary parts of the complex function $w = f(z)$ must fulfill in order that its derivative may have a unique value for any point z, regardless of the direction of the increment dz at this point. Equations 12 and 13 are known as the *Cauchy-Riemann partial differential equations* (*or condition equations*). Only those functions

$w = u + jv$ which satisfy these equations are henceforth to be called *functions of a complex variable.*

Practically all the common functions familiar to the reader are found to satisfy the conditions 12 and 13. The simplest of these functions is

$$w = z \qquad [14]$$

for which

$$u = x \quad \text{and} \quad v = y \qquad [15]$$

is obviously a function of a complex variable.

A constant times an integer power of z, namely

$$w = az^n \qquad [16]$$

is likewise seen to satisfy Eqs. 12 and 13. Hence it follows that any polynomial

$$w = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \qquad [17]$$

or any quotient of polynomials

$$w = \frac{a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0}{b_k z^k + b_{k-1} z^{k-1} + \cdots + b_1 z + b_0} \qquad [18]$$

are also functions of a complex variable. The familiar trigonometric, hyperbolic, and exponential functions as well as the logarithm, when regarded as functions of z, all satisfy Eqs. 12 and 13. Fractional powers of z and fractional powers of polynomials in z satisfy the conditions. It is, in fact, more difficult to find functions that do not satisfy the conditions 12 and 13 than it is to find those that do. A few exceptions are

$$w = |z| = \sqrt{x^2 + y^2} \qquad [19]$$

and

$$w = \bar{z} = x - jy \qquad [20]$$

but even these simple exceptions are rather peculiar and hardly worth bothering with anyway.

It is appropriate to point out that the fulfillment of the Cauchy-Riemann equations does not suffice for the existence of the derivative. The latter requires that the partial derivatives 12 and 13 be continuous functions of x and y in the vicinity of the point in question.

A point at which the function is not differentiable is called a *singularity.* If the function is differentiable everywhere within an arbitrarily small region in the vicinity of some point, it is there said to be *regular* or *analytic* (the term *holomorphic* is also used to describe this property). A region throughout which a function is analytic is spoken of as a region of *analyticity.*

2. A GRAPHICAL REPRESENTATION; CONFORMAL MAPPING

Since the complex function w as well as the complex variable z require two quantities (their real and imaginary parts) for their description, the graphical representation used for functions of a real variable is not available for the plotting of functions of a complex variable. Instead, the values of the variable z are plotted in one complex plane (the x,y-plane or z-plane), and the corresponding values of the function w are plotted in another complex plane (the u,v-plane or w-plane). A given point in the z-plane represents a complex value for the independent variable z which determines a value for the function w, and this value in turn determines a corresponding point in the w-plane. A continuous curve in the z-plane may be thought of as defining a set of points in this plane, and if the function $w = f(z)$ is continuous throughout this range of z-values, a corresponding continuous curve is thereby determined in the w-plane.

The construction of a family of curves throughout a given region in the z-plane makes it possible to map the behavior of the function over a corresponding region in the w-plane. For this purpose the curves drawn in the z-plane may, for example, be the sets of straight lines defined by x = constant and y = constant, the constants being chosen so that these orthogonal families of lines form a uniform grid. Alternatively, a set of circles concentric with the origin and the orthogonal set of radial lines may be drawn in the z-plane, or one may select any other sets of curves which appear best to serve the purpose in view of the particular function under consideration.

It is useful to note a very interesting property of functions of a complex variable which is made evident by such a graphical representation. This property follows directly from the fact that the derivative of the function is independent of the direction of the vector increment dz in the z-plane. Assuming for the moment that the increments are finite, the derivative is approximated by $\Delta w/\Delta z$. If the increment Δz is thought of as having a fixed magnitude but any desired angle, the fact that the quotient $\Delta w/\Delta z$ has the same complex value regardless of the angle of Δz means that variations in the angle of Δw are exactly equal to any assumed variations in the angle of Δz. In other words, since the angle of the complex quotient $\Delta w/\Delta z$ is independent of the angle of Δz, the *changes* in the angles of Δz and Δw must always be equal.

Any two increments $\Delta_1 z$ and $\Delta_2 z$ differing only in direction at a given point in the z-plane may be looked upon as a pair of path increments along any two curves which intersect at that point; and, similarly, the two corresponding increments $\Delta_1 w$ and $\Delta_2 w$ (whose magnitudes are alike because $\Delta_1 w/\Delta_1 z = \Delta_2 w/\Delta_2 z$) may be looked upon as a pair of path

increments along two curves intersecting at the corresponding point in the w-plane. Since the angle between the path increments $\Delta_1 z$ and $\Delta_2 z$ equals the angle between the corresponding path increments $\Delta_1 w$ and $\Delta_2 w$, any sets of curves drawn in the z-plane intersect at the same angles as the corresponding curves in the w-plane at all corresponding points. The process of mapping curves in the w-plane corresponding to any chosen curves in the z-plane, in other words, preserves the angular relationships between these curves at all corresponding points. For example, if two sets of curves drawn in the z-plane form orthogonal families, the corresponding sets of curves which are mapped in the w-plane by means of any function $w = f(z)$ (satisfying the conditions 12 and 13, of course) likewise form orthogonal families.

It should be observed that if the angle of the increment $\Delta_2 z$ is larger (or smaller) than the angle of $\Delta_1 z$, the angle of $\Delta_2 w$ is likewise larger (respectively smaller) than that of $\Delta_1 w$. In other words, the angular increment between $\Delta_1 w$ and $\Delta_2 w$ is equal to the angular increment between $\Delta_1 z$ and $\Delta_2 z$ not only in magnitude but also in *sense*. That is, if the rotation from $\Delta_1 z$ to $\Delta_2 z$ in the z-plane is, for example, counterclockwise, the rotation from $\Delta_1 w$ to $\Delta_2 w$ in the w-plane is also counterclockwise.

If $w = f(z)$ is a function satisfying the conditions 12 and 13, the function $\overline{w} = f(z)$, in which the bar indicates the conjugate value, evidently does not satisfy these conditions. Since, for a given Δz, the increments Δw and $\Delta \overline{w}$ have the same magnitudes but opposite angles, it is clear that the function $\overline{w} = f(z)$, in its mapping properties, preserves the angular relationships in magnitude but reverses their sense (as is the case with a picture and its mirror image). The mapping properties of both the functions $w = f(z)$ and $\overline{w} = f(z)$ are said to be *isogonal* (meaning that the magnitudes of angular relationships are preserved). In addition, the mapping property of the function $w = f(z)$, which preserves the sense as well as the magnitude of angular relationships, is described as being *conformal*.

As a consequence of the property of conformality, one may see that if a small region of a map in the z-plane (with numerous intersecting curves) is compared with the corresponding small region in the w-plane (with numerous corresponding intersecting curves), these two small mapped regions are found to be exact replicas of each other except for a factor of magnification (or diminution) equal to the magnitude of $\Delta w/\Delta z$ at the point where this region is located, and a rotation through the angle of $\Delta w/\Delta z$. This observation is strictly accurate, of course, only in the limit as the size of the entire region tends to zero, but for small regions of finite size, the two corresponding maps are very nearly alike in detailed form. The term " conformal " is thus seen to assume a clearer significance.

3. The inverse function

The corresponding maps in the w- and z-planes for a given function $w = f(z)$ place in evidence a mutual relationship between the two complex quantities w and z in the sense that either one may apparently be looked upon as the independent variable. In other words, the given function

$$w = f(z) = u(x,y) + jv(x,y) \tag{21}$$

may presumably be inverted to yield

$$z = \phi(w) = x(u,v) + jy(u,v) \tag{22}$$

at least over regions throughout which a one-to-one relationship exists between w and z. This thought may be investigated further through the consideration of the relations

$$dx = \frac{\partial x}{\partial u}\,du + \frac{\partial x}{\partial v}\,dv \tag{23}$$

$$dy = \frac{\partial y}{\partial u}\,du + \frac{\partial y}{\partial v}\,dv \tag{24}$$

which are the inverse of Eqs. 7 and 8.

Denoting the determinant of Eqs. 7 and 8 by D, and noting Eqs. 12 and 13, one sees that

$$D = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 = \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \tag{25}$$

But, again with the help of Eqs. 12 and 13, one has*

$$\frac{dw}{dz} = f'(z) = \frac{\partial u}{\partial x} + j\frac{\partial v}{\partial x} = \frac{\partial u}{\partial x} - j\frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} + j\frac{\partial v}{\partial x} \tag{26}$$

so that Eq. 25 yields

$$D = |f'(z)|^2 \tag{27}$$

Now Eqs. 7 and 8, on the one hand, and the pair of inverse Eqs. 23 and 24, on the other, must have inverse matrices; that is,

$$\begin{bmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{bmatrix}^{-1} \tag{28}$$

*The correctness of these relations should be clear from the fact that the value of the derivative is independent of the angle of the increment $dz = dx + j\,dy$. If this angle is zero then $dy = 0$, which means that u and v are differentiated *partially* with respect to x only.

from which it follows that

$$\frac{\partial x}{\partial u} = \frac{1}{D}\frac{\partial v}{\partial y}, \qquad \frac{\partial x}{\partial v} = -\frac{1}{D}\frac{\partial u}{\partial y} \tag{29}$$

$$\frac{\partial y}{\partial u} = -\frac{1}{D}\frac{\partial v}{\partial x}, \quad \frac{\partial y}{\partial v} = \frac{1}{D}\frac{\partial u}{\partial x} \tag{30}$$

Hence

$$\frac{\partial x}{\partial u} = \frac{\partial y}{\partial v} \tag{31}$$

$$\frac{\partial x}{\partial v} = -\frac{\partial y}{\partial u} \tag{32}$$

which are the Cauchy-Riemann equations pertaining to the inverse function 22.

With the help of these various relations, one may now write

$$\frac{dz}{dw} = \frac{\partial x}{\partial u} + j\frac{\partial y}{\partial u} = \frac{\dfrac{\partial v}{\partial y} - j\dfrac{\partial v}{\partial x}}{D} = \frac{1}{\dfrac{\partial v}{\partial y} + j\dfrac{\partial v}{\partial x}}$$

$$= \frac{1}{\dfrac{\partial u}{\partial x} + j\dfrac{\partial v}{\partial x}} = \frac{1}{f'(z)} = \left(\frac{dw}{dz}\right)^{-1} \tag{33}$$

According to Eqs. 31 and 32, the inverse function 22 is also a function of a complex variable, and Eq. 33 shows that the derivative of the inverse function is the reciprocal of that for the given function at corresponding values of w and z. In other words, the conformal maps for the function $w = f(z)$ may likewise be regarded as the maps for the inverse function $z = \phi(w)$.

A precautionary remark may be made here about difficulties of interpretation in dealing with multivalued functions. Although these matters are discussed in greater detail in subsequent articles it should be observed now that, in view of Eq. 33, the derivative of the inverse function evidently does not exist at a point where that of the given function $w = f(z)$ becomes zero. In the immediate vicinity of such points, the maps in the w- and z-planes for a given function are still uniquely related, although the preservation of angular relationships no longer holds (as is further discussed in Art. 14).

4. The z-plane and its associated complex sphere; the point at infinity

At various times it is expedient to consider the value or behavior of a function at infinity, that is, in the limit $z \to \infty$. Since any point in the complex z-plane which is infinitely remote from the origin is a point at infinity, it may seem as though infinity should be regarded as a vast region embracing an infinity of points. While admissible on purely logical grounds, this view is extremely awkward from a mathematical standpoint, since the behavior of a function "at infinity" would embrace its behavior at an infinite number of points.

The difficulty involved here is readily overcome, however, through introducing (by definition) a slightly altered conception of the complex plane. Thus it is perfectly admissible to think of this plane as being the surface of a sphere of infinite radius, or, for the sake of easing the mental strain produced by this conception, as a sphere with so large a radius that any finite region that may be considered appears for all practical purposes to be flat. If the origin of this "plane" is taken to be the south pole of the sphere, all points infinitely remote from the origin coincide at the north pole. Infinity is then no longer a vast region but becomes a single point. It is called *the* point at infinity.

In order to overcome the necessity of thinking of the z-plane as an enormous spherical surface, another artifice may be utilized which in some respects has certain advantages over the infinite sphere idea. The z-plane is visualized as a truly flat surface, with a sphere of arbitrary but finite radius resting upon its origin. The point of tangency between the sphere and the z-plane at its origin may be taken as the south pole of the sphere. The corresponding north pole is then perpendicularly above the origin.

A given finite point z_0 in the z-plane is now thought of as joined with the north pole of the sphere by a straight line, which intersects the surface of the sphere at one point other than the north pole. This geometrical construction, which is called *stereographic projection*, associates a point on the sphere with every point in the complex z-plane in a manner unique for all points except those infinitely remote from the origin. All these correspond to the north pole of the sphere. Thus, with the help of stereographic projection, infinity is again interpreted as a single point.

In all considerations regarding regions and paths in the complex plane, the corresponding ones on the surface of the sphere may be substituted. It can be shown geometrically that the process of stereographic projection preserves the magnitudes of angular relationships between intersecting curves in the plane and the corresponding ones on the spherical surface (the process is isogonal). Circles in the plane are circles on the sphere. In particular, any circle on the sphere which passes through the

north pole becomes a straight line (circle with infinite radius) in the plane. A great circle through the north pole (meridian) corresponds to a straight line drawn through the origin (radius vector) in the z-plane.

Such a sphere is referred to as the *complex sphere* associated with the z-plane. A similar sphere may evidently be associated with the w-plane in connection with the mapping of any function $w = f(z)$. Because of the isogonality of the process of stereographic projection, it follows that corresponding maps on the two spheres for a given function are conformal, so that the spherical surfaces may in all instances be used to replace the complex w- and z-planes. In this way the process of conformal mapping may readily be visualized over regions which include the point at infinity.

5. Alternative graphical and physical interpretations

A number of additional interesting properties of functions of a complex variable may be studied through identifying the complex plane with a cross-sectional plane associated with a physical system having longitudinal uniformity. This direction coincides with what is ordinarily designated as the z-axis of a rectangular co-ordinate system. A static field (electric, magnetic, or hydrodynamic) associated with such a supposed physical system has either a zero or a constant component in the longitudinal direction. This component is ignored. In the following discussion, and wherever the physical argument requires three-dimensional or space consideration, it is understood that a unit length in the longitudinal direction is implied. In other words, the field is regarded as a two-dimensional one, since its behavior is of interest only in the cross-sectional plane which is identified with the complex x,y-plane.

In such a system, a vector function $A(x,y)$ is assumed to represent the flow of some physical or fictitious fluid. In complex form

$$A = A_x + jA_y \qquad [34]$$

With reference to Fig. 1, S represents a closed (mathematical) boundary. If a differential length of this boundary is denoted by ds, the net flow outward through the boundary is given by the integral

$$\oint_S A_n \, ds \qquad [35]$$

in which A_n is the normal component (directed outward) of A. According to the geometry shown in the figure, and on the assumption that the integration is extended around the closed contour in the counterclockwise direction, this integral may be written

$$\oint_S A_n \, ds = \oint_S (A_x \, dy - A_y \, dx) \qquad [36]$$

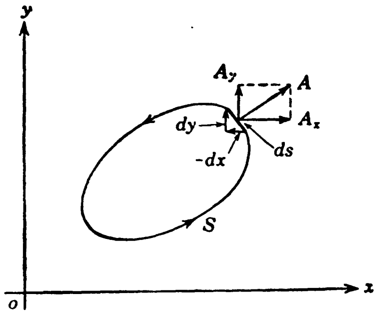


FIG. 1. Relevant to the integral of the normal component of a vector function around a closed contour.

From Gauss's law, it is recognized that this net flow outward may alternatively be calculated by integrating the divergence of A over the surface* enclosed by the boundary S, that is

$$\oint_S A_n\, ds = \int_{\text{enclosed surface}} \operatorname{div} A\, da \qquad [37]$$

But

$$\operatorname{div} A = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} \qquad [38]$$

and hence Eqs. 36 and 37 yield

$$\oint_S (A_x\, dy - A_y\, dx) = \int_{\text{enclosed surface}} \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y}\right) da \qquad [39]$$

If A_x and A_y are now identified respectively with the imaginary and the real parts of a function of a complex variable $w = f(z)$, then†

$$A_x = v(x,y) \qquad [40]$$

$$A_y = u(x,y) \qquad [41]$$

and the Cauchy-Riemann Eq. 13 shows that

$$\operatorname{div} A = 0 \qquad [42]$$

and hence that

$$\oint_S (A_x\, dy - A_y\, dx) = \oint_S (v\, dy - u\, dx) = 0 \qquad [43]$$

In order that the divergence of A with components defined by Eqs. 40 and 41 shall be zero throughout the surface enclosed by the boundary S, it is necessary, of course, that the Cauchy-Riemann equations hold throughout this region. This condition requires that the function $w = f(z)$ be regular throughout the region; otherwise its derivative does not exist at all points over which the surface integral in Eq. 37 or Eq. 39 extends.

Physically, Eq. 43 means that the field A is source-free throughout

*The reader may again be reminded that a unit of length in the longitudinal direction is implied so that this surface integration is actually equivalent to a volume integration.

†It should be observed that $w = A_y + jA_x$ and hence that $w = f(z)$ should not be confused with the vector function A, Eq. 34.

the enclosed region, as evidenced by the vanishing of the divergence of A for all points within the region. Hence it may be said that the imaginary and real parts respectively of a function of a complex variable which is regular throughout a given region in a z-plane may there be regarded as the x- and y-components of a source-free field.

If now the line integral of the vector function A is formed for the closed boundary shown in Fig. 1, with a counterclockwise direction of traversal, one has

$$\oint_S A \cdot ds = \oint_S (A_x\, dx + A_y\, dy) \qquad [44]$$

According to Stokes's law, it is recalled that

$$\oint_S A \cdot ds = \int_{\text{enclosed surface}} (\text{curl } A) \cdot da \qquad [45]$$

The curl of A, which is directed normal to the x,y-plane, is given by what would normally be regarded as the z-component, that is,

$$\text{curl}_z A = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \qquad [46]$$

Equations 44 and 45, therefore, yield

$$\oint_S (A_x\, dx + A_y\, dy) = \int_{\text{enclosed surface}} \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) da \qquad [47]$$

Again identifying A_x and A_y with v and u, respectively, according to Eqs. 40 and 41, and making use of the Cauchy-Riemann Eq. 12, one finds that

$$\text{curl } A = 0 \qquad [48]$$

If the function $w = f(z)$ is regular throughout the region enclosed by the curve S, Eq. 48 holds for all points within this region, so that

$$\oint_S (A_x\, dx + A_y\, dy) = \oint_S (v\, dx + u\, dy) = 0 \qquad [49]$$

Hence it may be said that the imaginary and real parts respectively of a function of a complex variable which is regular throughout a given region in the z-plane may there be regarded as the x- and y-components of a field which is not only source-free but also nonturbulent. If the enclosed region contains points at which the derivative of the function $w = f(z)$ does not exist, the relations 43 and 49 no longer hold. In view of the present discussion such singular points may be regarded as either sources or vortexes in which the origin of the field A, and hence that of the function $w = f(z)$, resides. This interpretation makes it clear that unless

the function w has singular points somewhere in the z-plane, it must reduce to a constant or to zero. In other words, the singularities of a function are its "life-giving" elements, and out of their nature and distribution alone does a function derive its individual properties and characteristics. This view leads to a useful method of classifying functions purely in terms of the nature and distribution of their singularities, which is briefly discussed later on.

The Cauchy-Riemann Eqs. 12 and 13 provide a further physical interpretation for the real and imaginary parts of a function of a complex variable. Thus if Eq. 12 is differentiated partially with respect to x and Eq. 13 with respect to y, the subsequent addition of the two equations yields

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{50}$$

On the other hand, if Eq. 12 is differentiated partially with respect to y and Eq. 13 with respect to x, the subsequent subtraction of the two equations gives

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \tag{51}$$

These results are recognized to have the form of Laplace's equation for the potential of a two-dimensional source-free field. Hence the real and imaginary parts of a function of a complex variable may be interpreted as scalar potential functions. As such they may be assumed to determine a pair of nonturbulent vector field functions. If these are denoted by A and B respectively, one may write

$$A_x = \frac{\partial u}{\partial x}, \qquad A_y = \frac{\partial u}{\partial y} \tag{52}$$

and

$$B_x = \frac{\partial v}{\partial x}, \qquad B_y = \frac{\partial v}{\partial y} \tag{53}$$

Because of the Cauchy-Riemann equations, these field components are related as expressed by

$$A_x = B_y, \qquad A_y = -B_x \tag{54}$$

and hence the scalar product of A and B vanishes; thus

$$A \cdot B = A_x B_x + A_y B_y = -A_x A_y + A_x A_y = 0 \tag{55}$$

In other words, the two nonturbulent fields defined by the potential functions u and v are orthogonal to each other.

Now it is further recalled (from the study of vector analysis) that the system of equipotential lines defined by the equations u = constant, and the flow lines for field A form orthogonal families of curves. The same is true of the equipotential lines defined by the equations v = constant and the flow lines for field B. Since the flow lines for field B are orthogonal to those for field A, it follows, therefore, that the equipotential lines defined by u = constant are orthogonal to those defined by v = constant. Hence the latter coincide with the flow lines for field A, and the former coincide with the flow lines for field B. This situation forms an alternative basis for the graphical representation of a function of a complex variable and for its physical interpretation.

Thus, instead of using the conformal maps in the w- and z-planes, one may study a given function of a complex variable graphically by plotting, in the z-plane alone, the systems of mutually orthogonal curves defined by the equations u = constant and v = constant. Throughout regions in which the given function $w = f(z)$ is regular, these have the character of the equipotential lines and flow lines of a source-free, nonturbulent field. Singular points again have the character of sources or vortexes, the nature and distribution of which determine the properties of the given function w. Some aspects of these physical interpretations are discussed further in Art. 22.

6. Integration; the Cauchy integral law

A certain orientation with regard to the question of differentiation having been gained, attention may now be directed toward the interpretation of the integral of a function of a complex variable. Here the two-dimensional character of the independent variable again injects some novel considerations at the outset. Thus if the integral

$$\int_{z_1}^{z_2} f(z)\, dz \qquad [56]$$

is formally regarded as representing the integral of a given function $w = f(z)$ between two particular values z_1 and z_2 of the independent variable z, the question of the choice to be made for the continuous sequence of values of z as it proceeds from the point z_1 to the point z_2 immediately arises. Such a continuous sequence of values evidently defines some path or curve joining the points z_1 and z_2 in the complex z-plane. Since any number of paths may evidently be chosen in the detailed process of evaluating the integral 56, the possibility exists that the value of this integral may not be unique.

This question is similar to that arising in connection with the discussion of the differentiation of complex functions, and it is again due to the two-

dimensional character of the independent variable. Again it is felt to be highly desirable that the value of the integral 56 should be unique, and if possible, that the Cauchy-Riemann equations, which insure the uniqueness of the derivative, should be sufficient to insure also the uniqueness of the integral 56 so that its value may be independent of the path of integration without need for the imposition of further conditions upon the function w.

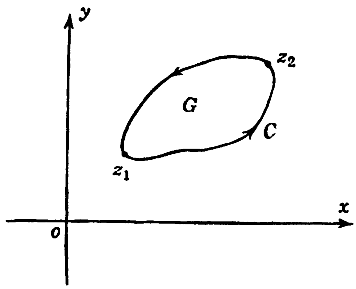


FIG. 2. Region of analyticity in the discussion of Cauchy's integral law.

In order to investigate this question it is expedient to consider the integral formed for a closed contour in the z-plane, as shown in Fig. 2. This so-called ***contour integral*** is written

$$\oint_C f(z)\, dz \qquad [57]$$

Here the contour C is assumed to be traversed in the counterclockwise direction, whence the enclosed region G is observed to lie on the left. Within the region G the function $w = f(z)$ is assumed to be regular. For all points within the region G, therefore, the function is differentiable, and the Cauchy-Riemann equations are fulfilled.

By substitution from Eqs. 2 and 3, the integral 57 becomes

$$\oint_C f(z)\, dz = \oint_C (u + jv)(dx + j\, dy)$$

$$= \oint_C (u\, dx - v\, dy) + j \oint_C (v\, dx + u\, dy) \qquad [58]$$

The closed contour C in Fig. 2 and the conventions regarding the direction of traversal are essentially the same as those shown in Fig. 1 for the closed boundary S; and since the function $w = f(z)$ is assumed to be differentiable at all points within the enclosed region, the results expressed by Eqs. 43 and 49 apply. Hence the important result follows that

$$\oint_C f(z)\, dz = 0 \qquad [59]$$

Since the points z_1 and z_2 appearing in the integral 56 may be thought of as any two points on the contour C, as indicated in Fig. 2, it follows from Eq. 59 that the integral 56 is independent of the path of integration or that a given path joining z_1 with z_2 may be changed at will without altering the value of the integral 56 so long as the path is not moved across a point at which the function w is singular. The latter restriction is

readily appreciated through noting that if the two portions of the contour C joining z_1 and z_2 in Fig. 2 are regarded as two variations of a given path between these points, the statement that the enclosed region G contains no singularities is seen to be equivalent to stating that none are encountered in the process of sweeping one of these paths across the region G into coincidence with the other.

The result 59 is known as the *Cauchy integral law*, which states in effect that the integral of a function of a complex variable between two given points in the complex z-plane has a unique value (subject, of course, to the restriction that any two chosen paths enclose a region in which the function is regular). The necessary conditions to insure this result are expressed by the Cauchy-Riemann equations which at the same time insure the uniqueness of the derivative of a given complex function.

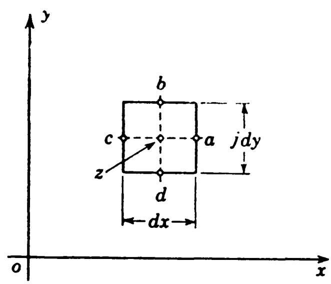


FIG. 3. An elementary closed path in the consideration of Cauchy's integral law.

This result is so important for all the subsequent discussions that it is well to consider in another way the relations leading to it. Figure 3 shows a differential rectangle in the z-plane with its center located at some point z. The midpoints of the sides of this rectangle are denoted by a, b, c, d. A given function has the value $w = f(z)$ at the point z and is there assumed to be regular. Except for differentials of higher order, its values at the points a, b, c, d are given by

$$f(a) = w + \frac{\partial f}{\partial x}\frac{dx}{2} \qquad [60]$$

$$f(b) = w + \frac{\partial f}{j\,\partial y}\frac{j\,dy}{2} \qquad [61]$$

$$f(c) = w - \frac{\partial f}{\partial x}\frac{dx}{2} \qquad [62]$$

$$f(d) = w - \frac{\partial f}{j\,\partial y}\frac{j\,dy}{2} \qquad [63]$$

Now since the function w satisfies the Cauchy-Riemann condition equations, its derivative is independent of the direction of the increment in the z-plane, so that

$$\frac{\partial f}{\partial x} = \frac{\partial f}{j\,\partial y} = \frac{df}{dz} \qquad [64]$$

and hence

$$f(a) = w + \frac{df}{dz}\frac{dx}{2} \tag{65}$$

$$f(b) = w + \frac{df}{dz}\frac{j\,dy}{2} \tag{66}$$

$$f(c) = w - \frac{df}{dz}\frac{dx}{2} \tag{67}$$

$$f(d) = w - \frac{df}{dz}\frac{j\,dy}{2} \tag{68}$$

The integral of the function w formed (in the counterclockwise direction) for the differential rectangle is given by the summation

$$\oint f(z)\,dz = f(a)j\,dy - f(b)\,dx - f(c)j\,dy + f(d)\,dx \tag{69}$$

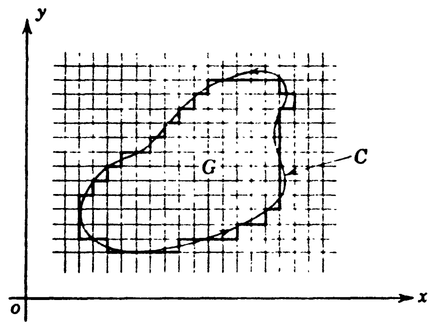


FIG. 4. Approximation of a contour by a rectangular step curve.

Substituting from Eqs. 65 to 68 shows that

$$\oint f(z)\,dz = 0 \tag{70}$$

As illustrated crudely in Fig. 4, a given closed contour C of finite size may be thought of as approximated by a rectangular step curve. The approximation becomes better and better as the size of the steps is made smaller and smaller. The integral around the closed contour C may be replaced by the sum of integrals around all the enclosed small rectangles (all taken in the counterclockwise direction) because the contributions from the sides of adjacent internal rectangles cancel, just as in the argument leading to the result known as Stokes's law in vector analysis. If the function $w = f(z)$ is differentiable at all points within the region

G, the result 70 is applicable to all the enclosed small rectangles, and thus the result 59 is again established. The Cauchy-Riemann equations together with the condition that w be differentiable at all points within the

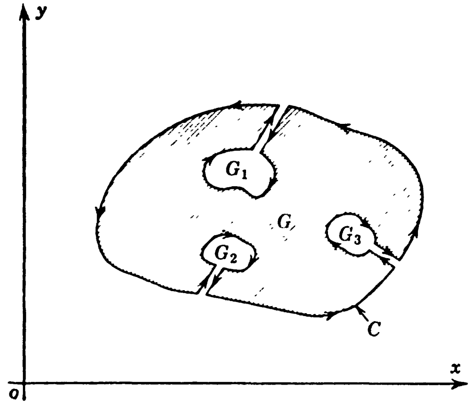


FIG. 5. A simply connected region.

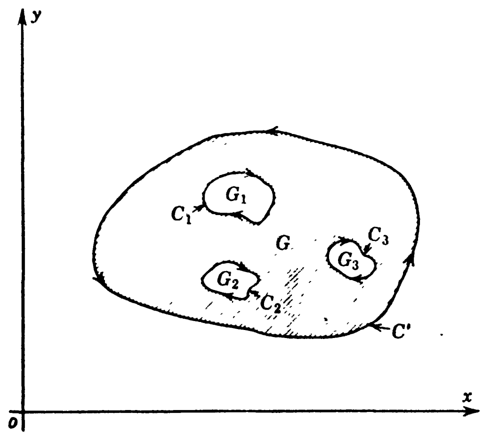


FIG. 6. A multiply connected region.

enclosed region are again found to be necessary and sufficient for the validity of Eq. 59.

The following remarks about the characteristics of the region enclosed by the contour C are necessary. This contour may conceivably have a form like that shown in Fig. 5, for which the enclosed region G is the shaded area. If the portions of the contour C leading to and from the smaller islandlike regions G_1, G_2, and G_3 within G are moved closer and

closer together until they finally become superimposed, the contour integral around C evidently becomes equivalent to the sum of four separate integrals evaluated respectively for the contours C', C_1, C_2, and C_3 with directions of traversal as indicated in Fig. 6.

So long as $w = f(z)$ is regular within the shaded region G, the integral law 59 still holds when evaluated for the contour C of Fig. 5 or for the equivalent set of contours C', C_1, C_2, and C_3 shown in Fig. 6. It may no longer hold, however, if applied only to the contour C' of Fig. 6. Actually the enclosed regions G_1, G_2, and G_3 may contain points at which the function $w = f(z)$ is not differentiable and therefore, the integrals pertaining to the separate contours C_1, C_2, and C_3 are not necessarily zero.

A region which has embedded in it one or more subregions $G_1, G_2, \cdots$ like the region G of Fig. 6 is said to be *multiply connected*. It is doubly connected if it contains one subregion, triply connected if it contains two subregions, etc. Unless the contour C in the integral 59 is interpreted in the manner illustrated in Figs. 5 and 6, the validity of Cauchy's integral law evidently requires that the region enclosed by the contour be *simply connected*.

7. Cauchy's integral formula

A given function is assumed to be regular within the region enclosed by the contour S shown in Fig. 7. It is also assumed to be regular at all points on the boundary S. These points are denoted by ζ, and the corresponding values of the function are expressed by

$$w = f(\zeta) \qquad [71]$$

A contour integral is now considered which has the form

$$\oint_S \frac{f(\zeta)\,d\zeta}{(\zeta - z)} \qquad [72]$$

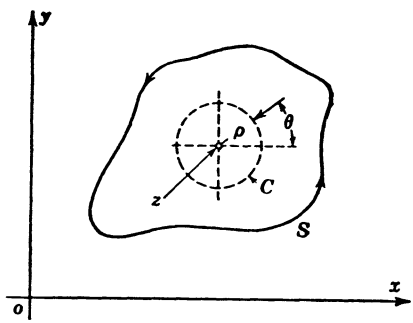


Fig. 7. The region of analyticity in the derivation of Cauchy's integral formula is the boundary S and the region enclosed by it.

Here z denotes some point within the enclosed region. If ζ is for the moment regarded as capable of assuming any value within the enclosed region as well as those on its boundary, it is observed that the integrand as a function of ζ is regular at all such points, with the exception of the point $\zeta = z$. At this point the integrand becomes infinite, and therefore Cauchy's integral law no longer applies; the value of the integral 72 is then not necessarily zero.

However, if the point z is surrounded by a circular contour C, and the integration is extended over a closed path which consists of the curve S and the circle C joined by a line which is traversed in both directions, as shown in Fig. 8, the result is evidently zero because the corresponding enclosed region is the shaded area where the integrand is regular. It follows, therefore, that the value of the integral 72 is equal to that of

$$\oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z)} \qquad [73]$$

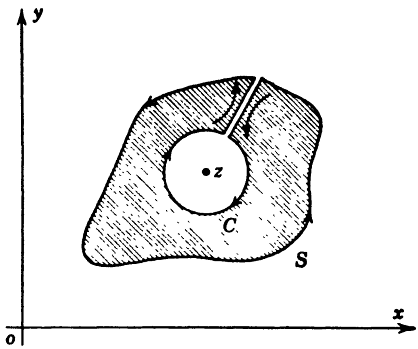


FIG. 8. The shaded region within which $\dfrac{f(\zeta)}{\zeta - z}$ is regular.

in which the circular contour C is traversed in the counterclockwise direction, and ζ now refers to points on this circular contour. The correctness of this statement is recognized from the fact that the difference between the integrals 72 and 73 is the integral around the composite boundary of Fig. 8, which has the value zero.

The radius ρ of the circle C about z is now to be thought of as being very small, so small in fact that throughout the process of carrying out the integration of 73 the value of $f(\zeta)$ differs from that of $f(z)$by a negligible amount. In view of the fact that $f(\zeta)$ is a continuous function within the region bounded by S, this is a permissible assumption. (It would obviously not be permissible if $f(\zeta)$ were discontinuous at the point z.) Now if $f(\zeta)$ in the integral 73 is replaced by $f(z)$, it may be taken outside the integral sign, because the integration is evaluated with respect to the variable ζ. Thus one has

$$\oint_S \frac{f(\zeta)\, d\zeta}{(\zeta - z)} = \oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z)} = f(z)\oint_C \frac{d\zeta}{(\zeta - z)} \qquad [74]$$

With reference to Fig. 7, the polar form of $(\zeta - z)$ is

$$(\zeta - z) = \rho e^{j\theta} \qquad [75]$$

and since ζ is constrained to lie on the circle C,

$$d\zeta = j\rho e^{j\theta}\, d\theta = j(\zeta - z)\, d\theta \tag{76}$$

so that

$$\frac{d\zeta}{(\zeta - z)} = j\, d\theta \tag{77}$$

Hence

$$\oint_C \frac{d\zeta}{(\zeta - z)} = j\int_0^{2\pi} d\theta = 2\pi j \tag{78}$$

and consequently Eq. 74 yields

$$\oint_S \frac{f(\zeta)\, d\zeta}{(\zeta - z)} = 2\pi j f(z) \tag{79}$$

or

$$f(z) = \frac{1}{2\pi j}\oint_S \frac{f(\zeta)\, d\zeta}{(\zeta - z)} \tag{80}$$

This result is known as *Cauchy's integral formula.* It enables one to calculate the value of a function of a complex variable at a point within a region in terms of its values on the boundary, provided the function is known to be analytic throughout the region inclusive of the boundary.

8. The existence of derivatives of any order

By means of Cauchy's integral formula, Eq. 80, it is possible to show that a function of a complex variable possesses derivatives of any order at a point where the function is regular. The familiar formula for the derivative of a function $w = f(z)$ reads

$$\frac{dw}{dz} = \underset{\Delta z \to 0}{\text{limit}}\left[\frac{f(z + \Delta z) - f(z)}{\Delta z}\right] \tag{81}$$

According to the formula 80,

$$f(z + \Delta z) = \frac{1}{2\pi j}\oint_S \frac{f(\zeta)\, d\zeta}{\zeta - (z + \Delta z)} \tag{82}$$

Substitution of Eqs. 80 and 82 into Eq. 81 gives

$$\frac{dw}{dz} = \underset{\Delta z \to 0}{\text{limit}}\left[\frac{1}{2\pi j}\oint_S \frac{f(\zeta)}{\Delta z}\left\{\frac{1}{\zeta - (z + \Delta z)} - \frac{1}{\zeta - z}\right\} d\zeta\right] \tag{83}$$

Now

$$\frac{1}{\zeta - (z + \Delta z)} - \frac{1}{\zeta - z} = \frac{\Delta z}{(\zeta - z)^2 - \Delta z\,(\zeta - z)} \tag{84}$$

so that the limit 83 yields

$$\frac{dw}{dz} = \frac{1}{2\pi j}\oint_S \frac{f(\zeta)\,d\zeta}{(\zeta - z)^2} \tag{85}$$

The second derivative, according to the form of Eq. 81, is expressed by

$$\frac{d^2w}{dz^2} = \operatorname*{limit}_{\Delta z \to 0}\left[\frac{f'(z + \Delta z) - f'(z)}{\Delta z}\right] \tag{86}$$

in which the prime denotes the first derivative. By Eq. 85

$$f'(z + \Delta z) = \frac{1}{2\pi j}\oint_S \frac{f(\zeta)\,d\zeta}{[\zeta - (z + \Delta z)]^2} \tag{87}$$

and Eq. 86 therefore gives

$$\frac{d^2w}{dz^2} = \operatorname*{limit}_{\Delta z \to 0}\left[\frac{1}{2\pi j}\oint_S \frac{f(\zeta)}{\Delta z}\left\{\frac{1}{[\zeta - (z + \Delta z)]^2} - \frac{1}{(\zeta - z)^2}\right\} d\zeta\right] \tag{88}$$

But

$$\frac{1}{[\zeta - (z + \Delta z)]^2} - \frac{1}{(\zeta - z)^2} = \frac{2(\zeta - z)\,\Delta z + \overline{\Delta z}^2}{(\zeta - z - \Delta z)^2(\zeta - z)^2} \tag{89}$$

so that the limit 88 is found to yield

$$\frac{d^2w}{dz^2} = \frac{2}{2\pi j}\oint_S \frac{f(\zeta)\,d\zeta}{(\zeta - z)^3} \tag{90}$$

Continuing in this way, one establishes the following formula for the nth derivative:

$$\frac{d^nw}{dz^n} = \frac{n!}{2\pi j}\oint_S \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{n+1}} \tag{91}$$

It may be concluded that the function $w = f(z)$, if regular within a given region, possesses derivatives of any order for all points of this region, and these derivatives may be obtained through differentiating under the integral sign in the Cauchy integral formula 80. The existence of but a single derivative thus implies the existence of all subsequent derivatives! The elegance of this remarkable result can hardly be overemphasized, nor is it out of place to call attention to the fact that a similar result does not apply to real functions.

The values of the derivatives of higher order are seen to be unique if this is true for the value of the first derivative of a given function. The first and all higher order derivatives of a function of a complex variable are themselves functions of a complex variable, and these derivative functions are analytic at any point where the given function is analytic.

9. Point sets and infinite series

Preliminary to the discussion of the expansion of functions in infinite series, which is given in the following article, it is helpful to point out a few fundamental principles regarding the convergence of such series. This subject may best be approached through considering an infinite set of points

$$s_1, \quad s_2, \quad s_3, \cdots s_k, \cdots \tag{92}$$

Visualize the quantities s_1, s_2, etc., as points in the complex plane. They may be scattered about completely at random, or the sequence of their values may exhibit a tendency to become confined within a more and more limited area or region as the index k (identifying individual values in the sequence) becomes larger and larger. In the latter case the sequence is said to approach a limit and the set of points is said to possess a *limit point.*

If one considers a number of concentric circles (with finite, nonzero radii) drawn about the limit point S in the complex plane in which the points s_1, s_2, $\cdots$ are indicated by dots, each of these circles, however small, contains an infinite number of points. This statement is the definition of a limit point. Considering a particular circle of radius ϵ, it is possible to state that some integer n exists such that all points s_k for $k > n$ lie inside this circle. For any given radius ϵ, the appropriate value of n must be sufficiently large to assure that none of the points in the unending sequence s_{n+1}, s_{n+2}, s_{n+3} $\cdots$ lie outside the circle (although some s_k for $k < n$ may still be inside); or, for a given integer n, the appropriate radius ϵ must be sufficiently large. As the value of n is chosen larger and larger the appropriate ϵ may be chosen smaller and smaller since the points become denser and denser in the more immediate vicinity of the limit point. The latter, for this reason, is sometimes also referred to as a *cluster point* or as a *point of condensation.*

It thus becomes clear that a limit s of the unending sequence s_1, s_2, s_3, $\cdots$ may be said to exist if

$$|S - s_k| < \epsilon, \quad \text{for all } k > n \tag{93}$$

in which ϵ is nonzero but may be chosen arbitrarily small, and n is a finite integer depending upon ϵ. If this condition is fulfilled, the sequence is said to *converge* to the limit S.

An important result known as the *Bolzano-Weierstrass theorem* follows from the definition of a limit point. This theorem states that if an unending sequence of points s_1, s_2, s_3, $\cdots$ is confined within a finite region, that region must contain at least one limit point. The truth of this statement may be seen through considering the region to be subdivided into a finite number of smaller ones, for example, squares. Since the number of points

is infinite and the number of squares finite, it is clear that at least one of the squares must contain an infinite number of points. This one may again be subdivided into smaller squares and the same reasoning repeated. Through continuing in this way one may state that within the original region it must be possible to find at least one nonzero but arbitrarily small subregion containing an infinite number of points, whence, according to the definition of a limit point, that subregion must contain one.

The convergence condition 93 may be expressed in an alternate form, known as *Cauchy's principle of convergence*, which reads

$$|s_{n+p} - s_n| < \epsilon, \quad \text{for a finite } n \text{ and all } p = 1, 2, \cdots \infty \tag{94}$$

According to this statement, a circle drawn about s_n with the finite radius ϵ contains all s_k for $k > n$; that is, it contains the unending sequence $s_{n+1}, s_{n+2}, s_{n+3}, \cdots$, and in view of the Bolzano-Weierstrass theorem it must, therefore, contain a limit point.

Consider now the infinite series

$$S = \sum_{n=1}^{\infty} u_n = u_1 + u_2 + u_3 + \cdots \tag{95}$$

and its *partial sums*

$$s_k = u_1 + u_2 + u_3 + \cdots + u_k \tag{96}$$

For $k = 1, 2, 3, \cdots$ these partial sums may be regarded as elements of the unending sequence $s_1, s_2, s_3, \cdots$ discussed above, and S as its limit. If this limit exists, the infinite series 95 is said to converge; if the limit does not exist, the series diverges. Cauchy's principle of convergence 94 may be expressed in the form

$$|u_{n+1} + u_{n+2} + \cdots + u_{n+p}| < \epsilon, \quad \text{for a finite } n \text{ and all } p = 1, 2, \cdots \infty \tag{97}$$

The quantity $|S - s_k|$ appearing in the condition 93 is the absolute value of the remainder of the series 95 after the kth term.

Beside the series 95 it is significant to consider the one expressed by

$$\sigma = \sum_{n=1}^{\infty} |u_n| = |u_1| + |u_2| + |u_3| + \cdots \tag{98}$$

in which the terms are the absolute values of corresponding ones in 95. Since the remainder

$$|u_{n+1} + u_{n+2} + u_{n+3} + \cdots| \tag{99}$$

of the series 95 is always smaller than the remainder

$$|u_{n+1}| + |u_{n+2}| + |u_{n+3}| + \cdots \tag{100}$$

of the series 98, it follows that 95 is surely convergent if 98 converges. In the latter event the series 95 is said to be *absolutely* convergent.

A sufficient test for the absolute convergence of the series 95 is expressed by the condition

$$\lim_{n \to \infty} |u_n|^{1/n} < 1 \tag{101}$$

To prove this statement one should first observe that no matter how the limit of $|u_n|^{1/n}$ is approached as n is allowed to become larger and larger, whether monotonically from above or from below or in an oscillatory fashion, the fact that the limit is less than unity enables one to state that there must always exist a finite value of n such that for $k > n$, $|u_k|^{1/k} \leqq \rho < 1$. Denoting the partial sums of the series 98 by σ_n, one may write

$$|\sigma_{n+p} - \sigma_n| = |u_{n+1}| + |u_{n+2}| + \cdots + |u_{n+p}| \leqq \rho^{n+1} + \rho^{n+2} + \cdots + \rho^{n+p} \tag{102}$$

or

$$|\sigma_{n+p} - \sigma_n| \leqq \rho^{n+1}(1 + \rho + \rho^2 + \cdots + \rho^{p-1}) = \rho^{n+1}\frac{(1-\rho^p)}{(1-\rho)} < \frac{\rho^{n+1}}{1-\rho} \tag{103}$$

If one chooses $\rho^{n+1}/(1 - \rho) = \epsilon$, then ϵ can be made arbitrarily small through the choice of a sufficiently large n. Therefore the condition 101 leads to the result

$$|\sigma_{n+p} - \sigma_n| < \epsilon, \quad \text{for a finite } n \text{ and all } p = 1, 2, \cdots \infty \tag{104}$$

which is Cauchy's condition 94 for the convergence of the series 98.

If $|u_n|^{1/n}$ approaches a limiting value $\rho > 1$ as n becomes larger and larger, the series 95 diverges since $|u_n| \to \rho^n > 1$ for large n, in violation of the obvious necessary convergence condition $u_n \to 0$ for $n \to \infty$.

Sometimes the test 101, for one reason or another, is not applicable and one must employ other means for examining the convergence of a series. A useful alternate method is the *d'Alembert ratio test* which is expressed in the statement: If in addition to $|u_n| \to 0$ for $n \to \infty$, one has

$$\left|\frac{u_{k+1}}{u_k}\right| < \rho < 1 \text{ for all } k > n \tag{105}$$

the series S, Eq. 95, is absolutely convergent.

To prove that this condition insures the fulfillment of the Cauchy condition 94, one may begin by observing that

$$\begin{aligned} |u_{n+1}| &< |u_n|\rho \\ |u_{n+2}| &< |u_{n+1}|\rho < |u_n|\rho^2 \\ |u_{n+3}| &< |u_{n+1}|\rho^2 < |u_n|\rho^3 \\ &\text{etc.} \end{aligned} \tag{106}$$

Hence

$$|\sigma_{n+p}-\sigma_n| = |u_{n+1}|+|u_{n+2}|+\cdots+|u_{n+p}|<|u_n|(\rho+\rho^2+\rho^3+\cdots+\rho^p) \tag{107}$$

or

$$|\sigma_{n+p} - \sigma_n| < |u_n| \times \frac{\rho - \rho^{p+1}}{1 - \rho} < |u_n| \frac{\rho}{1 - \rho} \tag{108}$$

Since $|u_n| \to 0$ for $n \to \infty$, one can always find a value of n beyond which $|u_n|\rho/(1 - \rho)$ is equal to or less than an arbitrarily small nonzero ϵ. Hence the Cauchy criterion is met.

Another useful method is the so-called *comparison test*, according to which the given series 95 is compared with another series

$$\mathfrak{S} = \sum_{n=1}^{\infty} v_n = v_1 + v_2 + v_3 + \cdots \tag{109}$$

Thus if $\mathfrak{S}$ is known to be absolutely convergent and $|u_n| < C|v_n|$, where C is any finite positive constant, the series S must be absolutely convergent. The series $\mathfrak{S}$ is referred to as the *dominant* of S (or is said to dominate the series S).

Trivial as this test may seem to most engineers, it is nevertheless very useful, particularly when both the preceding methods of testing for absolute convergence fail. An example is furnished by the series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^r} = 1 + \frac{1}{2^r} + \frac{1}{3^r} + \frac{1}{4^r} + \cdots \tag{110}$$

Here

$$|u_n|^{1/n} = \left(\frac{1}{n}\right)^{r/n} \to 1 \text{ with } n \to \infty \text{ for all finite } r \tag{111}$$

This last result may be seen by observing that

$$\ln\left(\frac{1}{n}\right)^{r/n} = r\left(\frac{1}{n}\ln\frac{1}{n}\right) \tag{112}$$

Since the quantity in the parenthesis becomes zero for $n \to \infty$ the result 111 follows at once. It is clear that the condition 101 is not met for any finite nonzero r, and yet one cannot say with certainty that the series diverges, since inequality 101 is a sufficient, but not necessary, condition for convergence.

The criterion 105 for the d'Alembert ratio test likewise leads to an indecisive result, namely, one has

$$\left|\frac{u_{n+1}}{u_n}\right| = \left(\frac{n}{n+1}\right)r \to 1 \quad \text{for } n \to \infty \tag{113}$$

The convergence of the series 110 may be investigated through first determining another series which dominates it. To this end one writes

$$S = \sum_{n=1}^{\infty} \frac{1}{n^r} = 1 + \left(\frac{1}{2^r} + \frac{1}{3^r}\right) + \left(\frac{1}{4^r} + \frac{1}{5^r} + \frac{1}{6^r} + \frac{1}{7^r}\right) +$$

$$\left(\frac{1}{8^r} + \frac{1}{9^r} + \frac{1}{10^r} + \frac{1}{12^r} + \frac{1}{13^r} + \frac{1}{14^r} + \frac{1}{15^r} + \frac{1}{16^r}\right) + \cdots \quad [114]$$

Now

$$\left(\frac{1}{2^r} + \frac{1}{3^r}\right) < \frac{1}{2^r} + \frac{1}{2^r} = \frac{2}{2^r} = 2^{1-r}$$

$$\left(\frac{1}{4^r} + \frac{1}{5^r} + \frac{1}{6^r} + \frac{1}{7^r}\right) < \frac{4}{4^r} = (2^{1-r})^2$$

$$\left(\frac{1}{8^r} + \frac{1}{9^r} + \cdots + \frac{1}{16^r}\right) < \frac{8}{8^r} = (2^{1-r})^3 \quad [115]$$

and so forth. It is clear, therefore, that the series

$$\mathfrak{S} = \sum_{n=0}^{\infty} (2^{1-r})^n = 1 + (2^{1-r}) + (2^{1-r})^2 + \cdots \quad [116]$$

dominates the series 110 or 114. The right-hand side of Eq. 116 has the form of the power series

$$S = \sum_{n=0}^{\infty} z^n = 1 + z + z^2 + z^3 + \cdots \quad [117]$$

for which $|u_n|^{1/n} = |z|$, so that the condition 101 reads

$$|z| < 1 \quad [118]$$

With regard to the series 116 the corresponding condition leads to

$$|2^{1-r}| < 1 \quad \text{or} \quad r > 1 \quad [119]$$

When this condition is fulfilled, the series 110 converges.

For $r = 1$ this series reads

$$S = \sum_{n=0}^{\infty} \frac{1}{n} = 1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \cdots \quad [120]$$

and

$$|s_{n+p} - s_n| = \left|\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{n+p}\right| > \frac{p}{n+p} \quad [121]$$

For no finite n, however large, can this quantity be less than an arbitrarily small (nonzero) ϵ for all $p = 1, 2, \cdots \infty$, since for $p > n$ its value approaches unity. Therefore the series is divergent for $r = 1$; and it is

certainly divergent for smaller values of r since the terms in the corresponding expression 121 then are all larger than they are for $r = 1$.

Absolutely convergent series have the important fundamental property that their values are unaltered through a rearrangement of the terms. A convergent series that is not absolutely convergent is referred to as being *conditionally* convergent. The value of such a series can very definitely be altered through a rearrangement of its terms. In fact Riemann has shown that a conditionally convergent series can be made to have any finite value for its sum through an appropriate grouping of the terms.

For example, the series

$$S = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \tfrac{1}{5} - + \cdots \qquad [122]$$

is evidently not absolutely convergent since the sum of the absolute values of its terms is the series 120 which is divergent. It is not difficult to see by inspection, however, that the series does approach a finite limit in an oscillatory fashion. This limit may be computed to any desired accuracy, taking the terms in their given order. One thus finds $S = 0.69315 \cdots$ which, by other means, may be shown to be the decimal fraction approximation to $\ln 2$. If the terms are now rearranged as follows

$$S = 1 + \tfrac{1}{3} - \tfrac{1}{2} + \tfrac{1}{5} + \tfrac{1}{7} - \tfrac{1}{4} + \tfrac{1}{9} + \tfrac{1}{11} - \tfrac{1}{6} + \cdots \qquad [123]$$

and the limit is again computed, one finds

$$S = 1.03972 \cdots = 1.5 \ln 2$$

When a given series is not absolutely convergent, its possible conditional convergence may be investigated in the following way. Let the series be written in the form

$$S = \sum_{n=1}^{\infty} a_n v_n \qquad [124]$$

in which a_n and v_n are any two parts into which the typical term of the series is separable (one may need to make several trials at this step in order to find an appropriate separation). Defining the quantities

$$\begin{aligned} \mathfrak{S}_1 &= a_1 \\ \mathfrak{S}_2 &= a_1 + a_2 \\ \mathfrak{S}_3 &= a_1 + a_2 + a_3 \\ &\cdots\cdots\cdots\cdots \\ \mathfrak{S}_n &= a_1 + a_2 + \cdots + a_n \end{aligned} \qquad [125]$$

one may rewrite the series 124 as

$$S = \mathfrak{S}_1(v_1 - v_2) + \mathfrak{S}_2(v_2 - v_3) + \mathfrak{S}_3(v_3 - v_4) + \cdots \qquad [126]$$

with the partial sums

$$S_n = \mathfrak{S}_1(v_1 - v_2) + \mathfrak{S}_2(v_2 - v_3) + \cdots + \mathfrak{S}_{n-1}(v_{n-1} - v_n) + \mathfrak{S}_n v_n \qquad [127]$$

Thus one finds

$$\begin{aligned} |S_{n+p} - S_n| &= |-\mathfrak{S}_n v_n + \mathfrak{S}_n(v_n - v_{n+1}) + \cdots \\ &\quad + \mathfrak{S}_{n+p-1}(v_{n+p-1} - v_{n+p}) + \mathfrak{S}_{n+p}v_{n+p}| \\ &< |\mathfrak{S}_n| \cdot |v_n| + |\mathfrak{S}_n| \cdot |v_n - v_{n+1}| + \cdots \\ &\quad + |\mathfrak{S}_{n+p-1}| \cdot |v_{n+p-1} - v_{n+p}| + |\mathfrak{S}_{n+p}| \cdot |v_{n+p}| \end{aligned} \qquad [128]$$

Now if

$$|\mathfrak{S}_k| = |a_1 + a_2 + a_3 + \cdots + a_k| < A \quad \text{for all } k \qquad [129]$$

with A finite,

$$\begin{aligned} |S_{n+p} - S_n| < A(|v_n| + |v_{n+p}| + (|v_n - v_{n+1}| + |v_{n+1} - v_{n+2}| \\ + \cdots + |v_{n+p-1} - v_{n+p}|)) \end{aligned} \qquad [130]$$

Next, if

$$\sigma = \sum_1^\infty |v_n - v_{n+1}| \text{ is convergent} \qquad [131]$$

so that

$$|\sigma_{n+p} - \sigma_n| = |v_n - v_{n+1}| + |v_{n+1} - v_{n+2}| + \cdots + |v_{n+p-1} - v_{n+p}| < \epsilon_1 \qquad [132]$$

for a finite n and all $p = 1, 2, \cdots \infty$, and if further

$$|v_{n+p}| < \epsilon_2 \text{ for all } p = 0, 1, 2, \cdots \qquad [133]$$

the condition 130 becomes

$$|S_{n+p} - S_n| < A(\epsilon_1 + 2\epsilon_2) = \epsilon \qquad [134]$$

for a finite n and all $p = 1, 2, \cdots \infty$. The series 124 is then convergent, and the conditions 129, 131, 133 constitute a test (known as the Dedekind test for conditional convergence) which may be used to reveal this fact.

As an illustration the method may be applied to the series 122. Here one may let

$$a_n = (-1)^{n-1}, \qquad v_n = \frac{1}{n} \qquad [135]$$

Then $\mathfrak{S}_n = 1$ or zero and hence remains bounded for all values of n (condition 129). Terms in the series 131 have the form

$$|v_n - v_{n+1}| = \left|\frac{1}{n} - \frac{1}{n+1}\right| = \frac{1}{n(n+1)} \qquad [136]$$

This typical term is smaller than $1/n^2$, and since the series with $1/n^2$ as its typical term is convergent, the comparison test reveals the convergence of the series 131 in the present example. Finally the condition 133 is evidently met since v_n in 135 approaches zero monotonically with increasing n. Thus the conditional convergence of the series 122 is proved.

In the majority of problems in which infinite series appear, the terms of the series are functions of some independent variable, and its sum is likewise regarded as a function of this variable. For example, in the power series 117 the quantity z may be assumed to have any complex value. It does not necessarily follow, however, that the infinite series represents a function of this complex variable since values of the series are not related to values of the variable unless the series is convergent.

If for all points of a given region in the z-plane the series is convergent, it is said to be *uniformly* convergent in this region, and the series may then be regarded as there representing a function of the complex variable z. As the condition 118 shows, the power series 117 is convergent for all points lying inside the unit circle about the origin. This is its region of absolute and uniform convergence. If the power series is written

$$S = \sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots \qquad [137]$$

then the convergence condition 101 becomes

$$\operatorname*{limit}_{n\to\infty} |a_n|^{1/n}|z| < 1 \qquad [138]$$

or

$$|z| < \frac{1}{\operatorname*{limit}_{n\to\infty} |a_n|^{1/n}} \qquad [139]$$

This series converges absolutely and uniformly within a circle about the origin having the radius

$$r = \frac{1}{\operatorname*{limit}_{n\to\infty} |a_n|^{1/n}} \qquad [140]$$

This circle is called the *convergence circle* (also *circle of absolute convergence*) of the power series 137.

For points outside the convergence circle the series diverges, for there the inequality in the condition 138 is reversed. The series may still converge for some points on the circle, but for at least one such point the series must be divergent otherwise the radius of the convergence circle could be chosen larger than the value given by Eq. 140, and this conclusion is in conflict with the condition 138.

The statements just made with regard to the power series are unaltered

if the variable z is replaced by $(z - z_0)$ except that the convergence circle is then centered at the point $z = z_0$.

To say that an infinite series converges uniformly at some point $z = z_1$, is equivalent to saying that the series converges not only at this point but also for all points within a circle about z_1 with a nonzero but arbitrarily small radius. When this condition is met the term-by-term derivative of the series yields a resultant convergent series that correctly represents the derivative of the function defined by the given series, and an analogous statement may be made with regard to term-by-term integration. Within a region of uniform convergence one may, therefore, carry out the differentiation or integration of a function defined by an infinite series through applying the identical process to each term and summing afterward. The resulting series, however, may or may not possess the same region of uniform convergence.

Power series have the interesting property that their term-by-term derivative or integral yields resulting series with the same region of uniform convergence. This fact may easily be proved. The series obtained through differentiation and integration of 137 are respectively

$$S^{(1)} = \sum_{n=0}^{\infty} n a_n z^{n-1} \quad \text{and} \quad S^{(-1)} = \sum_{n=0}^{\infty} \frac{a_n}{n+1} z^{n+1} \qquad [141]$$

The condition 101 applied to these series reads

$$\lim_{n\to\infty} |na_n|^{1/n}|z|^{1-(1/n)} < 1 \quad \text{and} \quad \lim_{n\to\infty} \left|\frac{a_n}{n+1}\right|^{1/n} |z|^{1+(1/n)} < 1 \quad [142]$$

Since

$$\lim_{n\to\infty} |n|^{1/n} = \lim_{n\to\infty} \left|\frac{1}{n+1}\right|^{1/n} = 1 \qquad [143]$$

(as may be seen through considering the logarithms of the expressions subjected to the limiting process $n \to \infty$) it becomes clear that the conditions 142 coincide with the condition 138, and hence that the series 141 have the same convergence circle as the series 137.

These matters pertaining to uniform convergence of series may be illustrated through considering the straight-forward expansion of some simple algebraic functions. For example, if one divides $(1 - z)$ into unity by continuing the process of long division, there results

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + z^4 + \cdots \qquad [144]$$

which is the power series 117. For values of z within the unit circle this series identically replaces the function $1/(1 - z)$, whereas for values of z outside this circle the series and the function $1/(1 - z)$ have entirely

different values. Near the point $z = 1$ both the series and the function have values that increase without limit, whereas at the point $z = -1$, Eq. 144 yields

$$\tfrac{1}{2} = 1 - 1 + 1 - 1 + 1 - 1 + - \cdots \qquad [145]$$

The series is evidently divergent at this point, for it fails to approach a limit. Hence there is no reason to become puzzled about this result since the function and the series may be identified only where the latter converges.

However, in view of the fact that conditionally convergent series, through a particular arrangement of their terms, may be made to yield any desired sum, one is led to inquire whether a divergent series like the one in Eq. 145 may nevertheless be summable through the application of some special process.

This question, in the past, interested a number of mathematicians — Cesàro, Hölder, Euler, Abel, Borel, etc. — and the summation procedures developed by them are connoted by the letters C, H, E, A, B, etc., respectively. Thus if, for example, a particular series is summable through applying the Cesàro procedure, one abbreviates this statement by saying that the series is C-summable. This particular process of summation is now discussed in some detail.

With reference to the series 95 and its partial sums 96, consider the sequence

$$\begin{aligned}
S_1 &= s_1 = u_1 \\
S_2 &= \frac{s_1 + s_2}{2} = u_1 + \frac{u_2}{2} \\
S_3 &= \frac{s_1 + s_2 + s_3}{3} = u_1 + \frac{2}{3}u_2 + \frac{1}{3}u_3 \qquad [146] \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
S_n &= \frac{s_1 + s_2 + \cdots + s_n}{n} = u_1 + \left(1 - \frac{1}{n}\right)u_2 + \left(1 - \frac{2}{n}\right)u_3 + \cdots + \frac{1}{n}u_n
\end{aligned}$$

which is referred to as an *arithmetic mean sequence* of the first order (if this sequence is subjected to the same process the result is said to be of *second* order, etc.). A limit of the sequence 146 exists so long as s_n remains finite for unlimited values of n. It is not necessary that the terms of the series 95 have the property $u_n \to 0$ for $n \to \infty$, but, if they do, one is led to the result

$$\lim_{n\to\infty} \left\{u_1 + \left(1 - \frac{1}{n}\right)u_2 + \left(1 - \frac{2}{n}\right)u_3 + \cdots + \frac{1}{n}u_n\right\} = \lim_{n\to\infty} (s_n) \qquad [147]$$

That is, the new sequence then apparently has the same limit as the

sequence of partial sums s_n, a conclusion that should, however, be accepted with obvious reservations.

For the series in Eq. 145 the partial sums are

$$s_1 = 1,\ s_2 = 0,\ s_3 = 1,\ s_4 = 0, \cdots \qquad [148]$$

from which it is clear that

$$\operatorname*{limit}_{n\to\infty} (S_n) = \operatorname*{limit}_{n\to\infty} \left\{\frac{s_1 + s_2 + \cdots + s_n}{n}\right\} = \frac{1}{2} \qquad [149]$$

Thus the left-hand member of Eq. 145 is interpreted as the Cesàro sum of the divergent series on the right. This type of summation will be used in connection with Fourier series in the following chapter.

10. Taylor's and Maclaurin's series

The point of departure in the present argument is Eq. 79, which is repeated below:

$$\oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z)} = 2\pi j f(z) \qquad [150]$$

Here the contour integral is assumed to be evaluated for a circular path C with its center at some point z_0. Writing

$$\frac{1}{\zeta - z} = \frac{1}{(\zeta - z_0) - (z - z_0)} \qquad [151]$$

one obtains, by the process of long division, the series

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} + \frac{z - z_0}{(\zeta - z_0)^2} + \frac{(z - z_0)^2}{(\zeta - z_0)^3} + \cdots \qquad [152]$$

which, according to the theory of infinite series discussed in the previous article, is known to be uniformly convergent for

$$\left|\frac{z - z_0}{\zeta - z_0}\right| < 1 \qquad [153]$$

Since the center of the circular contour C lies at the points z_0, and ζ lies upon the circle, this condition for uniform convergence is fulfilled for all points z *within* the circle. Because of its uniform convergence, the series representation for $1/(\zeta - z)$ may be substituted into the integral Eq. 150. and the integration carried out term by term. This process gives

$$2\pi j f(z) = \oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)} + (z - z_0) \oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^2} + (z - z_0)^2 \oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^3} + \cdots \qquad [154]$$

A series representation for the function $w = f(z)$ is thus obtained which reads

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \cdots \qquad [155]$$

with

$$a_n = \frac{1}{2\pi j}\oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)^{n+1}} \qquad [156]$$

Equation 155 is recognized as Taylor's series representation for the function $f(z)$ in the vicinity of the point z_0, and Eq. 156 yields the coefficients for this expansion. Since the validity of Eq. 150 requires that the function $f(z)$ be regular throughout the region enclosed by the contour C, and the series 152 or 154 is uniformly convergent only for points within this circular contour (because of the condition 153), it follows that the Taylor series 155 converges uniformly only within a circle about z_0 whose contour reaches to the nearest singularity of the function $w = f(z)$.

It is useful to note that, by means of Eq. 91, the relation for the coefficients of the Taylor series expressed by Eq. 156 may be written in the alternate form

$$a_n = \frac{1}{n!}\left(\frac{d^n f}{dz^n}\right)_{z=z_0} \qquad [157]$$

which is recognized to agree with the familiar form used for the expansion of functions of a real variable. The Taylor series is thus found to be available in unaltered form to functions of a complex variable. The Maclaurin series evidently applies also to functions of a complex variable, since it is identical with the Taylor series for the special case $z_0 = 0$.

If the Taylor series, Eq. 155, is differentiated term by term, it is observed, according to the formula 157, tha the coefficients in the resulting series are those for the expansion of the function df/dz about the point $z = z_0$, and hence that this result represents the Taylor series for the derivative of $f(z)$. The truth of this statement may be shown through starting with Eq. 150, differentiating under the integral sign (as permitted according to the conclusions of Art. 8), and substituting the square of the series 152 for the quantity $1/(\zeta - z)^2$, obtaining in place of Eq. 155 a series for the function df/dz.

One may conclude from this reasoning that the term-by-term differentiation of the Taylor series is permitted and yields the Taylor series representation for the derivative of the given function. Moreover, the resulting series has the same region of convergence as that for the given function, as may be seen from the discussion in the previous article or from the fact (brought out in the closing sentence of Art. 8) that the

derivative of a function of a complex variable is again a function of a complex variable and possesses the same region of analyticity. From an obvious continuation of the same argument one sees that these conclusions apply to derivatives of any order.

Collateral to these considerations is the fact that the Taylor expansion of a given function is unique. To demonstrate this fact, assume that a series representation having the form of Eq. 155 is given, and that this series is somehow known to represent the function $f(z)$ within an arbitrarily small region in the immediate vicinity of the point $z = z_0$. The given series must be *the* Taylor expansion for $f(z)$ about the point $z = z_0$, because the behavior of the function in this vicinity completely determines its successive derivatives there. That is, differentiating the given series and substituting the results into Eq. 157 demonstrates that the coefficients are identical with those of a Taylor expansion about the same point.

11. The principle of analytic continuation

One of the most interesting things about the theory underlying the Taylor expansion is that the mere knowledge of a function within an arbitrarily small region in the vicinity of some point z_0 completely determines that function at all other points within the convergence circle about that point. The purpose of the present article is to call attention to the even more remarkable fact that a function of a complex variable is determined throughout the *entire* z-plane* from a knowledge of its properties within an arbitrarily small region of analyticity.†

The first step in the process of carrying out such a determination is to write down the Taylor series about the point z_0 where the function and its successive derivatives are known. Some other point z'_0 within the convergence circle for this Taylor series may then be chosen as the center about which a new Taylor series is determined. This choice can always be made, because the original series may be used for calculating the values

*This statement is subject to some restriction when the function in question possesses what is known as a natural boundary, every point of which is a singularity. In such an example, either the function does not exist beyond this boundary or its behavior there is governed by an entirely separate definition, in which case it is perhaps more appropriate to say that one is dealing not with a single function but actually with two separate functions. In order for the reader to appreciate that these ideas are not merely of an academic nature, his attention is called to the fact that natural boundaries of the sort referred to here do occur in practical problems. For example, the functions representing the fields in the cross-sectional plane of a wave guide possess as their natural boundary the walls enclosing these fields.

†It is not even necessary to know the function at *all* points within an arbitrarily small region about z_0; it is sufficient to know the values of the function, in the interior of the region, for all points of a finite but arbitrarily small line segment, or in a set of discrete points which have a limit point.

of the successive derivatives of the function in the point z'_0. Unless every point on the original convergence circle is singular, the point z'_0 can always be so chosen that the new convergence circle about z'_0 encloses a portion of the region lying outside the original circle. Since the two circles partially overlap, the two Taylor series have a certain convergence region in common where the function is determined by either series. The second series, however, enables one to calculate values of the function at points within a circumscribed region lying beyond the boundaries which limit the representation of the function by means of the first series. One speaks of this process as an *analytic continuation* of the function into the newly acquired region.

By properly choosing a third point z''_0 within this newly acquired region and using the second series for the calculation of the successive derivatives of the function in this point, one obtains a third Taylor series whose convergence circle encloses a portion of the z-plane which lies beyond the boundaries given by either of the first two convergence circles. The third Taylor series, therefore, represents a further continuation of the same function.

In order to obtain such a further continuation it is, of course, not necessary that the third point z''_0 be located outside the first convergence circle, since the selection of this point may be regarded merely as a revision in the choice of the second point z'_0. It should be easy to appreciate that this procedure may be continued in a variety of ways so as to obtain numerous overlapping convergence circles and corresponding series representations by means of which the function is ultimately determined within any desired region of the z-plane except for the singular points of the function.

The process of carrying out an analytic continuation in this manner and of obtaining a succession of partially overlapping convergence circles which extend the known region for the function into a continuously expanding portion of the z-plane may be regarded as a process of successively acquiring access to additional area in the z-plane after the fashion that a harvester, cutting grain with a scythe, successively acquires additional stubble ground by executing a continuous series of semicircular slices. If for the moment one imagines this harvester to be doing a rather unsystematic job, it is conceivable that he may cut a swath or path which circles about a portion of the grain field and returns so as to overlap itself. If the process of analytic continuation is carried out in this manner, one expects, in the overlapping portion, to regain the same values of the function as were obtained previously. This, however, may or may not be the case, for the function may be multivalued, and in returning to the original portion of the z-plane one may find oneself located on a different leaf of the Riemann surface* characterizing that multivalued

*These matters are discussed in further detail in Arts. 17 and 18.

function. In fact it is possible to find, after returning to the original region, that a point at which the function was observed to be regular when this region was first encountered is now a singularity of the function.

The significent matter here is that the process of analytic continuation is applicable whether the function is single-valued or not, and that even a multivalued function with all its manifold characteristics and peculiarities is completely determined from a knowledge of that function over an arbitrarily small region of analyticity on a single leaf of its Riemann surface.

It follows also from these considerations that any two functions of a complex variable whose values coincide over an arbitrarily small region of analyticity, or for all points of a finite but arbitrarily small line segment, or for a set of discrete points having a limit point, must have identical values throughout their common region of analyticity and hence must there be identical. This statement is known as the *identity theorem* or *uniqueness theorem* for analytic functions.

If the common region of analyticity possesses a natural boundary, nothing is implied regarding the behavior of the functions within other possible regions of analyticity. The theorem likewise does not imply that the behaviors of the two functions are identical at isolated singularities located within their common region of analyticity, but these matters are practically irrelevant inasmuch as a function is usable only where it is analytic, and no difficulty can arise from assuming (if this be desirable) that the functions are identical everywhere else.

The arbitrarily small line segment over which the function is initially known may be a portion of the real or imaginary axis in the z-plane. A real function of a real variable may be regarded as a function of a complex variable whose independent and dependent values both happen to be on the real axis. If it is possible to continue such a function into the complex domain, that continuation is unique and is immediately obtained by the simple expedient of replacing the real independent variable x by the complex variable $z = x + jy$. The truth of this statement, expressing a property of functions known as their *permanence of form*, follows directly from the identity theorem, since the given real function and its continuation obviously coincide for points on the real axis.

12. Singular points and the Laurent expansion

Since the singularities of a function of a complex variable are, so to speak, the mainsprings upon which its very existence depends, it is usually of chief interest in the study of a given function that characteristics of the function be investigated in the immediate vicinity of these singular points. For this purpose the Taylor series is of little service, because the vicinities of singular points are the very regions where its convergence fails. Consequently it is of considerable importance to search for a type

of expansion whose sphere of usefulness is centered about a singular point, that is, an expansion which places the character of a given singularity in evidence.

A few preliminary remarks concerning singularities will be appropriate preceding the detailed discussion of how such an expansion is found. Quite generally one must admit the possibility that a given function may be singular (nondifferentiable or nonanalytic) not only at certain discrete points but also at *all* points comprising a finite region in the z-plane. If the latter is the case, and a particular point within this region is singled out, it is not possible to discover an immediately surrounding or neighboring space, however small, in which the function is analytic. In other words, the singular points within such a region are infinitely dense.

For the considerations of the present article, such singular regions must be ruled out. Indeed, it is not possible to discover any kind of series expansion which can represent a function in the vicinity of a point located within a region of this sort. From a practical point of view this is hardly discouraging, however, inasmuch as functions which possess such singular regions are seldom encountered in engineering work.* The present discussion, then, will apply only to singularities for which it *is* possible to discover an immediately surrounding region in which no other singular points lie. These are called *isolated* singularities.

With reference to Fig. 9, z_0 represents a point at which a function $w = f(z)$ has an isolated singularity. About this point as a center are drawn a small circle c and a larger one C. The given function may have other singularities within the smaller circle or outside the larger one, but it is assumed to be analytic and single-valued at all points such as the point z inside the annular space between the two circles. Cauchy's integral formula, Eq. 79, is, therefore, applicable to the composite contour consisting of the two circles and the path joining them, traversed continuously as indicated in Fig. 9. Since the path joining the circles is traversed in both directions, the resulting form of Eq. 79 maybe written

$$2\pi j f(z) = \oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z)} + \oint_c \frac{f(\zeta)\, d\zeta}{(\zeta - z)} \tag{158}$$

in which the larger circle is traversed in the counterclockwise direction, the smaller one is traversed in the clockwise direction, and ζ refers to points on either of the two circles.

For the integral over the large circle, the series 152 for $1/(\zeta - z)$

*It may be pointed out in this connection that in terms of the analogy to physical flow fields (as discussed in Art. 5), the continuously distributed sources or vortexes (spatial distributions of electric charge or current densities, for example) constitute just such singular regions. However, the solution of practical problems dealing with these situations fortunately does not require the series representation of functions within these regions.

converges uniformly because the condition 153 is fulfilled. For the integral over the small circle, this statement is no longer true, because the point

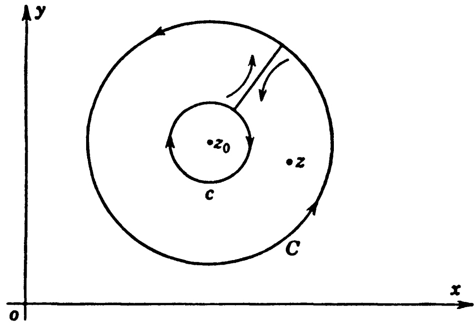


FIG. 9. The composite contour to which Cauchy's integral formula is applied in the derivation of the Laurent expansion.

z lies outside the small circle and hence $|z - z_0| > |\zeta - z_0|$. Here Eq. 151 must be replaced by

$$\frac{1}{\zeta - z} = \frac{-1}{(z - z_0) - (\zeta - z_0)} \qquad [159]$$

whence a process of long division yields the series

$$\frac{1}{\zeta - z} = -\frac{1}{z - z_0} - \frac{(\zeta - z_0)}{(z - z_0)^2} - \frac{(\zeta - z_0)^2}{(z - z_0)^3} - \cdots \qquad [160]$$

which is uniformly convergent for

$$\left|\frac{z - z_0}{\zeta - z_0}\right| > 1 \qquad [161]$$

It thus becomes clear that for $1/(\zeta - z)$ one may substitute the series 152 into the first integral in Eq. 158 and the series 160 into the second integral, and carry out both integrations term by term. Equation 158 then becomes

$$2\pi j f(z) = \begin{cases} \oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)} + (z - z_0) \oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)^2} \\ \qquad + (z - z_0)^2 \oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)^3} + \cdots \\ -(z - z_0)^{-1} \oint_c \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)^0} - (z - z_0)^{-2} \oint_c \frac{f(\zeta)\,d\zeta}{(\zeta - z_0)^{-1}} \\ \qquad -(z - z_0)^{-3} \oint_c \frac{f(\zeta)\,d\xi}{(\zeta - z_0)^{-2}} - \cdots \end{cases} \qquad [162]$$

This result may be written in the form

$$f(z) = b_0 + b_1(z - z_0) + b_2(z - z_0)^2 + b_3(z - z_0)^3 + \cdots$$
$$+ b_{-1}(z - z_0)^{-1} + b_{-2}(z - z_0)^{-2} + b_{-3}(z - z_0)^{-3} + \cdots \quad [163]$$

in which

$$b_n = \frac{1}{2\pi j}\oint_C \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^{n+1}} \quad \text{for } n = 0, 1, 2, \cdots \quad [164]$$

and

$$b_n = \frac{-1}{2\pi j}\oint_c \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^{n+1}} \quad \text{for } n = -1, -2, -3, \cdots \quad [165]$$

Equation 163 is the desired series representation for the function $w = f(z)$. It is called the *Laurent* series. From the derivation just given it is clear that the series converges uniformly within the annular region between the two circles shown in Fig. 9 where the function w is analytic. Thus the radius of the outer convergence circle extends from z_0 to the nearest singularity beyond the circumference of the smaller circle, and the radius of the inner convergence circle extends from z_0 to the farthest singularity inside the larger circle. If the singularity at z_0 is the only one inside the larger circle, the radius of the smaller circle may become vanishingly small, and the Laurent series is then seen to represent the given function in the immediate vicinity of the singularity at z_0, that is, to represent an expansion of $w = f(z)$ about this singularity as a center.

In connection with the formulas 164 and 165 for the coefficients of the Laurent series it should be observed that the direction of traversal about the large circle C in the integral 164 is counterclockwise, whereas that about the small circle c in the integral 165 is clockwise. If the latter direction of traversal is reversed, the algebraic sign in Eq. 165 changes from minus to plus. The formulas 164 and 165 then differ only in that the first is evaluated for the circumference of the larger circle and the second is evaluated for the circumference of the smaller one. Since the function $w = f(z)$ is analytic in the region between the two circles, the values of the integrals 164 and 165 are the same for any closed contour within the annular region or coincident with either circle. Hence the formulas 164 and 165 may be replaced by a single one which reads

$$b_n = \frac{1}{2\pi j}\oint_S \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^{n+1}} \quad [166]$$

in which S is any closed contour within the annular region or coincident with either circle, ζ refers to points on this contour, and the latter is traversed in the counterclockwise direction. Since the function $w = f(z)$

is singular in the point z_0, Eq. 91 does not apply for $z = z_0$ and hence there exists no alternative differential formula for the coefficients b_n, similar to Eq. 157 for the coefficients of the Taylor series. This circumstance is a practical disadvantage because integration is, as a rule, more difficult than differentiation. Consequently, when a Laurent expansion is to be found, the coefficients are determined, wherever possible, by other expedients. Further discussion of this point is given later on.

When the form of the Laurent series, Eq. 163 is contrasted with that of the Taylor series, Eq. 155, the principal difference is observed to lie in the fact that the Laurent series contains both descending as well as ascending powers of the variable $(z - z_0)$, whereas the Taylor series contains only the ascending powers. The portion of the Laurent series involving the ascending powers only is called the *ascending part* of the series, and the portion involving the descending powers is called the ***descending part*** or ***principal part***. It is the principal part of the Laurent series which places in evidence the singularity of the function $w = f(z)$ at the point z_0.

The ascending and descending parts of the Laurent expansion may be written respectively as

$$f_1(z) = \sum_{n=0}^{\infty} b_n(z - z_0)^n \qquad [167]$$

and

$$f_2(z) = \sum_{n=-1}^{-\infty} b_n(z - z_0)^n \qquad [168]$$

whence Eq. 163 for the Laurent series becomes

$$f(z) = f_1(z) + f_2(z) \qquad [169]$$

Because the series 152 (which leads to the ascending part f_1) converges uniformly for all z-values within the larger circle and the series 160 (which leads to the descending part f_2) converges uniformly for all z-values outside the smaller circle (Fig. 9), it follows that the series 167 converges everywhere within the larger circle whereas the series 168 converges everywhere outside the smaller one. Both series converge within the annular region between the two circles (this is the common portion of the two separate regions of convergence), and this, therefore, is the region of convergence for the sum of the two series 167 and 168, which is the Laurent expansion.

The function $f_1(z)$ is analytic everywhere within the larger circle, and the series 167 is its Taylor expansion about the point $z = z_0$. The function $f_2(z)$ is analytic everywhere outside the smaller circle. To interpret the series 168 for $f_2(z)$, it is helpful to consider for the moment a change of

variable which amounts to replacing $(z - z_0)$ by $w = 1/(z - z_0)$ which amounts to interchanging the roles of points z_0 and ∞. The series 168 then becomes one involving only ascending powers of the new variable w and represents a Taylor expansion of the function $f_2(z)$ about the point $w = 0$ (which corresponds to a Taylor expansion about the point $z = \infty$ according to the change of variable considered). It is helpful in this reasoning to think in terms of the complex sphere rather than the complex plane, for then the points $z = z_0$ and $z = \infty$ are simply two points on the sphere and the interchange of the parts which they play is easier to visualize. One is thus led to recognize that the descending series 168 may be regarded as a Taylor expansion of the function $f_2(z)$ about the point $z = \infty$.

According to this interpretation the Laurent series represents the given function by means of two Taylor series, one about the point $z = z_0$ (this is the ascending series for f_1) and one about the point $z = \infty$ (this is the descending series for f_2). The function $f_1(z)$ contains only those singularities of $f(z)$ which lie outside the larger circle about z_0; $f_2(z)$ contains only those singularities of $f(z)$ which lie within the smaller circle. This circle may alternatively be regarded as one which is drawn about the point $z = \infty$ as a center, whence the region " within " the circle (actually the region outside the smaller circle) becomes the region of analyticity for the function $f_2(z)$. The change of variable indicated by $(z - z_0) \rightarrow 1/(z - z_0)$ evidently interchanges not only the points $z = z_0$ and $z = \infty$ but also the roles played by the functions $f_1(z)$ and $f_2(z)$ and their series representations as given by Eqs. 167 and 168.

From the uniqueness theorem discussed in the preceding article it follows that the Laurent expansion is unique, since two series representations having identical regions of convergence and yielding identical values for all points within this region must be identical. This fact has practical significance, particularly with regard to the Laurent expansion, because it means that if a representation having the form of Eq. 163 is found for a given function in the vicinity of one of its singularities, this representation must be *the* Laurent series for that region irrespective of the method by which it is determined. In other words, the coefficients need not be calculated by means of the formula 166, but may be found by any other expedient which proves to be most effective in the given circumstances.

13. Kinds of Singularities and the Classification of Functions in Terms of Them

If the smaller circle c of Fig. 9 encloses no other singularities except the isolated one at the point $z = z_0$, the principal part (Eq. 168) of the Laurent series (Eq. 163) characterizes the nature of that singularity. If

the principal part contains a finite number of terms, the singularity is referred to as a *pole*. The principal part may then be written

$$h(z) = b_{-1}(z - z_0)^{-1} + b_{-2}(z - z_0)^{-2} + \cdots + b_{-s}(z - z_0)^{-s} \quad [170]$$

in which the highest negative exponent s is called the *multiplicity* or *order* of the pole. A pole of the first order is also called a *simple* pole.

The principal part $h(z)$ may have an infinite number of terms, in which case the function $f(z)$ is said to have an *essential singularity* at the point $z = z_0$. At a pole the function $f(z)$ becomes infinite, but in the vicinity of an essential singularity the function may assume *any* assigned value depending upon the manner in which this singularity is approached.* The function $w = e^z$, for example, has an essential singularity at the point $z = \infty$†, and the function $w = e^{1/\sin z}$ has essential singularities at the points $z = 0, \pi, 2\pi, \cdots$. For the function

$$f(z) = e^z = e^{x+jy} = e^x (\cos y + j \sin y) = u + jv \quad [171]$$

one may recognize this peculiarity by noting that

$$e^x = \sqrt{u^2 + v^2} \quad [172]$$

and

$$\tan y = \frac{v}{u} \quad [173]$$

Consider an arbitrary choice of values u and v. It is then possible to allow z to approach infinity along a path, parallel to the y-axis, designated by $x = \ln \sqrt{u^2 + v^2}$, since this value of x satisfies Eq. 172. Equation 173 may also be satisfied along this path for an infinite number of values of y which tend to infinity. Also, it approaches infinity along the negative real axis, e^z becomes zero, and if z approaches infinity along the positive real axis, e^z becomes infinite.

Not only is a pole a milder form of singularity, but the behavior of the function in its vicinity is a very definite one. For a pole of multiplicity s, multiplication of $f(z)$ by the factor $(z - z_0)^s$ yields a function which is regular in the point $z = z_0$. This circumstance may be regarded as a test whereby an ordinary pole may be distinguished from any other kind of singularity.

Other kinds of singularities, particularly those found in multivalued functions (branch points), are discussed in a subsequent article. In the meantime it is useful to point out how certain types of single-valued func-

*Theorem of Casorati-Weierstrass.

†This characteristic of the function e^z may be recognized from the fact that the Taylor series, which in this case converges uniformly in the entire z-plane, has an infinite number of terms. Note also the discussion of entire functions immediately following.

tions may be classified according to the poles or essential singularities which they possess. Once more it is emphasized that the singularities of a function are the mainsprings of its existence. Without singularities of any kind, an analytic function reduces to a constant.

In this classification one may begin with that type of function which is singular only in the point at infinity. Such a function is regular in the entire finite z-plane. It is called an *entire* function or also an *integral* function, and it may be denoted by $I(z)$. There are two kinds of entire functions; for one of these the singularity at infinity is an ordinary pole, and for the other it is an essential singularity. The first of these functions is more particularly referred to as an entire *rational;* the second, as an entire *transcendental* function.

Since the entire function is regular in the entire finite z-plane, it possesses a Taylor series representation for all finite points, which converges uniformly within the entire z-plane. If the function is transcendental, such a Taylor series contains an infinite number of terms. The functions $\sin z$, $\cos z$, and e^z are common examples. An entire rational function, on the other hand, possesses a Taylor series representation having a *finite* number of terms, the highest power of $(z - z_0)$ being equal to the order of the pole at infinity. The entire rational function, therefore, is an ordinary (finite) polynomial; that is,

$$I(z) = P(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots + a_n(z - z_0)^n \qquad [174]$$

in which n is the order of the pole at infinity. For $n = 0$, the function is also regular at infinity, and in this case reduces to the constant a_0.

A second important class of functions are those referred to as *meromorphic*. They may be defined as given by the ratio of two entire functions; thus

$$M(z) = \frac{I_1(z)}{I_2(z)} \qquad [175]$$

Since $I_2(z)$ is a finite or an infinite polynomial, it may become zero at a finite or at an infinite number of points in the z-plane. If $I_2(z)$ is thought of as factored in terms of its roots (these are called the *zeros* of I_2), it becomes clear that, at these points, $M(z)$ has ordinary poles whose orders equal the multiplicities of the roots of $I_2(z)$. The factored form of $I_2(z)$ is its finite or infinite product representation, and this form places the poles of the function $M(z)$ in evidence.

At the point infinity the meromorphic function has an essential singularity if either or both of the entire functions I_1 and I_2 are transcendental. $M(z)$ is then said to be transcendental also, but at no finite points in the z-plane can this function have singularities other than ordinary poles.

The function $w = \tan z$ is a common example of a meromorphic function which is also transcendental. It has an infinite number of simple poles in the z-plane and in addition has an essential singularity at infinity.

When both I_1 and I_2 are rational (given by finite polynomials), the resultant meromorphic function is also rational. At the point infinity, the function then has at most an ordinary pole. In other words, a meromorphic function whose singularity at infinity is an ordinary pole, is a *rational* function. Rational functions, then, are such having *no* other singularities except poles. Inasmuch as the representation of them is given by

$$R(z) = \frac{P_1(z)}{P_2(z)} \tag{176}$$

in which P_1 and P_2 are finite polynomials, it follows that a rational function has a finite number of poles.

14. Zeros and saddle points or points of stagnation

If in the Taylor series representation for the function $w = f(z)$ as given by Eq. 155, the constant term a_0 is zero but a_1 is not zero, the function is said to have a *simple* zero in the point $z = z_0$. For the immediate vicinity of this point, that is, for $(z - z_0) < 1$, the function is approximately represented by the term $a_1(z - z_0)$ alone. The reciprocal function, $1/f(z)$, has a simple pole in this same point, for its representation in this vicinity is approximately given by $1/a_1(z - z_0)$.

If both a_0 and a_1 are zero and a_2 is different from zero, the function is said to have a zero of the *second order* in the point $z = z_0$. For the immediate vicinity of this point the function then is approximately represented by the term $a_2(z - z_0)^2$. The reciprocal function similarly is approximately represented by $1/a_2(z - z_0)^2$ and has a pole of second order in this point.

In general a zero is said to be of the order s if the reciprocal function has a pole of order s in the given point. This is the case if the first nonzero coefficient in the Taylor series expansion for the function is a_s. According to the formula 157 for the Taylor coefficients, this condition results if the function and its first $s - 1$ derivatives all vanish at the point $z = z_0$.

It is also possible that the first $s - 1$ derivatives of the function are zero at the point $z = z_0$ but that the function itself is not zero there. That is to say, all the coefficients a_1, a_2, and so forth up to and including a_{s-1} are zero, but a_0 is not zero. In this case the function $f(z) - a_0$, or $w - a_0$, has a zero of the order s in the point $z = z_0$, but the function w obviously is not zero there. Except for the additive constant a_0, the behavior of the function w in this vicinity is, however, clearly the same as though this

point were an sth order zero. The terminology used to refer to such points must be so chosen, however, as to distinguish them from zeros. For reasons of physical interpretation, to be discussed in the following paragraphs, they are referred to as *saddle points* or also as *points of stagnation.**

For the immediate vicinity of a saddle point of the order $s - 1$, one has†

$$w = u + jv \cong a_0 + a_s(z - z_0)^s \tag{177}$$

The point is a zero of order s if $a_0 = 0$. The following detailed discussion, in which the constant a_0 is dropped, applies to either zeros or saddle points. It is convenient to write

$$(z - z_0) = re^{j\phi} \tag{178}$$

whence

$$w = u + jv \cong a_s r^s e^{js\phi} = a_s r^s(\cos s\phi + j \sin s\phi) \tag{179}$$

Equating real and imaginary parts gives

$$u \cong a_s r^s \cos s\phi \tag{180}$$

$$v \cong a_s r^s \sin s\phi \tag{181}$$

For the graphical representation, in the z-plane, of the loci for u = constant and v = constant, according to the discussion given in Art. 5, the relations 180 and 181 are helpful in showing the character of such loci in the vicinity of the point z_0.

These loci are shown in Fig. 10 for the cases $s = 1$, $s = 2$, and $s = 3$. The corresponding sketches in Fig. 11 show how the algebraic signs of the quantities u and v change in the vicinity of the point $z = z_0$. The pictures in Fig. 10 may be regarded as depicting the direction of the gradient (u = constant) and the lines of constant altitude (v = constant) in a mountainous terrain. The picture for $s = 2$, for example, is then seen to represent the vicinity of a point which is simultaneously the top of a ridge and the bottom of a valley, that is, a mountain pass where a valley crosses a ridge. The terrain in such a region evidently has the shape of a

*The German terminology (of which these are translations) is "Sattelpunkt" or "Staupunkt." Alternatively the term "Kreuzungspunkt" is also used.

†The convention of designating the order of a saddle point as being $s - 1$ when a zero having the same properties is referred to as being of the order s arises from the fact that the inverse function has a branch point where the given function has a saddle point, whereas the reciprocal function (not to be confused with the inverse function) has a pole where the given function has a zero. Just as the order of a pole receives the same designation as that of the zero of the corresponding reciprocal function, so the order of a saddle point receives the same designation as that of the branch point of the inverse function. These matters are discussed in greater detail at the end of Art. 18 after the method of dealing with multivalued functions is presented.

saddle. This fact accounts for the appropriateness of the term *saddle point*.

For $s = 3$ the terrain has the shape of a saddle which might be designed for a three-legged person, or one may say that it is a region where three ridges and three valleys meet in a common point.

Using a hydrodynamic analogy, one may regard the curves for u = constant as representing the direction of fluid flow, and the curves for v = constant as designating the orthogonal set of contours along which the gradient is zero. The fluid is streaming symmetrically toward and away from the point z_0. At this point the fluid is stagnant, thus

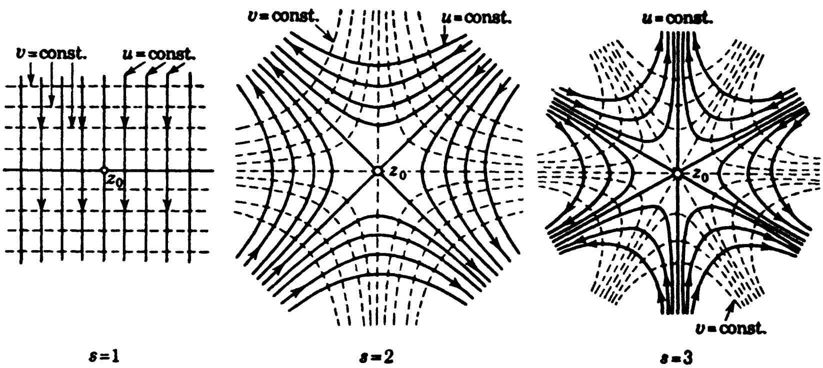


FIG. 10. Loci of constant real and imaginary parts in the vicinity of saddle points of various orders.

suggesting the term *point of stagnation* as an alternative designation.

The pictures show, moreover, that if the order of a zero is greater than unity (or that of a saddle point is greater than zero), the orthogonality of the contours for u = constant and v = constant fails in the point $z = z_0$, since the respective curves there intersect at an angle of $\pi/2s$ radians. This failure is, however, only apparent inasmuch as the contours actually do not pass through the point but are bent sharply at it. Fluid is deflected at the point z_0 instead of flowing through it.

From the standpoint of conformal mapping there is also an apparent failure in the angular relationships at a zero order greater than one (or at a saddle point of nonzero order). If, in addition to the polar representation of $(z - z_0)$ given by Eq. 178, one also writes

$$w - a_0 = \rho e^{j\phi} \qquad [182]$$

in which a_0 may or may not be zero, Eq. 177 shows that in the immediate vicinity of the point $z = z_0$ one has

$$\rho = a_s r^s \qquad [183]$$

and

$$\theta = s\phi \qquad [184]$$

Since $(z - z_0)$ may be regarded as a small path increment radiating from the point z_0 in the z-plane, and w may similarly be regarded as the

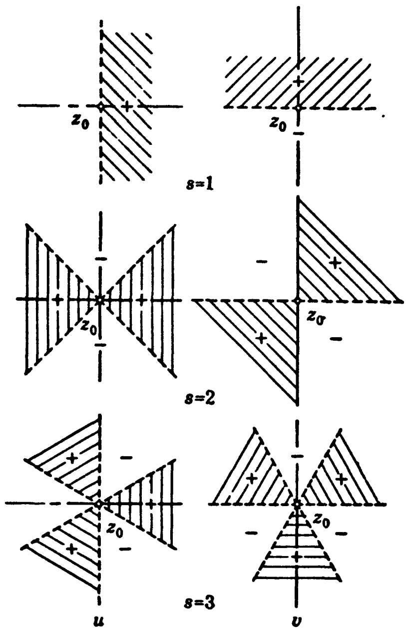


FIG. 11. Algebraic signs of real and imaginary parts in the vicinity of saddle points of various orders.

corresponding small path increment radiating from the point a_0 in the w-plane, one observes, according to Eq. 184, that if the increment $(z - z_0)$ is rotated through an angle $\Delta\phi$, the corresponding increment w rotates, not through the same angle (as the conformality ordinarily requires), but through an angle s times as large.

This apparent failure in the preservation of angular relationships is clarified by the recognition that the inverse function $z = \psi(w)$ is multivalued and that the point $z = z_0$ is a branch point of the order $s - 1$ for this inverse function. The discussion of conformal mapping in Art. 2 points out that such a pair of maps in the w- and z-planes is a graphical representation not only for the given function but also for the inverse of it. Consequently one must recognize that although the function given by Eq. 182 is single-valued, the multivalued character of the inverse function cannot be ignored.

Although the detailed properties of multivalued functions is left for discussion in Arts. 17 and 18, it is appropriate to point out here how one may see that the failure in the preservation of angular relationships is only an apparent one. If the point z_0 in the z-plane is enclosed by a sufficiently small circle c, the corresponding locus in the w-plane is very nearly a small circle c' surrounding the point w_0 (which corresponds to z_0). As a point traverses an arc on the circle c, the corresponding point on c' traverses an arc s times as great. In Fig. 12, a and b are two small radial line segments emanating from the circle c, and a' and b' are the corresponding segments in the w-plane. The angle θ between a' and b' is s

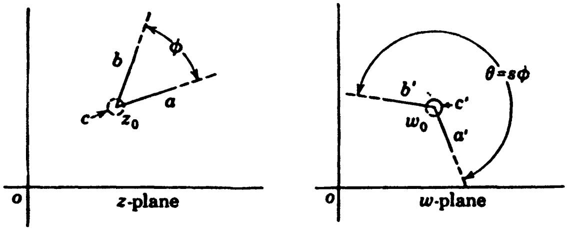


FIG. 12. Apparent failure in the preservation of angular relationships at a saddle point.

times that between a and b. If the radii of the small circles c and c' are allowed to become still smaller, the line segments appear to radiate from the points z_0 and w_0, and one is led to conclude that the preservation of angular relationships has failed because one's attention is focused upon the angles ϕ and θ rather than upon the angles between the line segments and the circular arcs, which remain equal to 90 degrees. However, if one mentally visualizes the situation in the limit as though the small circles were still there, it becomes clear that what has happened is not a failure in the preservation of angular relationships but rather is the result of a peculiar and somewhat misleading behavior of the given function in the vicinity of the point $z = z_0$. This view is borne out by the fact that the Cauchy-Riemann equations, guaranteeing the uniqueness of the derivative, still hold in this point.

15. The evaluation of contour integrals; Cauchy's residue theorem

If the formula given by Eq. 166 for the coefficients of the Laurent expansion is written for the integer value $n = -1$, it reads

$$\oint_S f(z)\,dz = 2\pi j b_{-1} \tag{185}$$

The contour S is in the present discussion assumed to enclose a region within which the function $f(z)$ has but one singularity at the point z_0. This situation is indicated in Fig. 13. If the coefficient b_{-1} can be determined in some way (for example, as described subsequently in this article), Eq. 185 represents a means for evaluating the contour integral for the function $f(z)$ extended around a given closed boundary S.

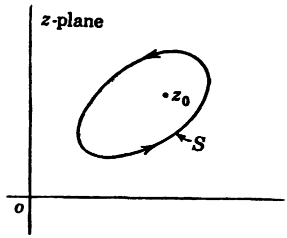


FIG. 13. The integral of a function about S is determined by the value of the residue of the pole at z_0.

According to the Cauchy integral law, as discussed in Art. 6, the value of this contour integral is zero if the given function $f(z)$ is regular at all points enclosed by the contour. The present result substantiates this fact, for if $f(z)$ is regular also in the point z_0, the Laurent expansion about this point has no principal part (it becomes identical with the Taylor expansion) and hence $b_{-1} = 0$. The present result may be said to represent a completion of the Cauchy integral law in the sense that it yields the value of the contour integral whether the function is regular within the enclosed region or not, and hence it contains the integral law as a special case.

An interesting alternative way of obtaining this same result is to begin by assuming that the contour integral is given and that an evaluation of it is sought. Since the singularity at z_0 is the only one within the region, the contour S may be replaced by a circle C about z_0 according to reasoning similar to that used in Art. 7 in replacing the integral 72 by the integral 73. In other words, the path of integration may be deformed or contracted so long as no part of it is allowed to sweep over a singular point. Thus

$$\oint_S f(z)\,dz = \oint_C f(z)\,dz \qquad [186]$$

Now $f(z)$ may be replaced by its Laurent expansion which reads

$$f(z) = \sum_{n=-\infty}^{\infty} b_n(z - z_0)^n \qquad [187]$$

Because of the uniform convergence of this series, the integration may be carried out term by term, giving

$$\oint_S f(z)\,dz = \sum_{n=-\infty}^{\infty} b_n \oint_C (z - z_0)^n\,dz \qquad [188]$$

If the radius of the circle C about z_0 is denoted by ρ, then

$$(z - z_0) = \rho e^{j\phi} \qquad [189]$$

and

$$dz = j\rho e^{j\phi}\,d\phi \qquad [190]$$

Equation 188 becomes

$$\oint_S f(z)\,dz = \sum_{n=-\infty}^{\infty} jb_n\rho^{n+1}\int_0^{2\pi} e^{j(n+1)\phi}\,d\phi \qquad [191]$$

But

$$\int_0^{2\pi} e^{j(n+1)\phi}\,d\phi = \begin{cases} 2\pi & \text{for } n = -1 \\ 0 & \text{for } n \neq -1 \end{cases} \qquad [192]$$

Hence

$$\oint_S f(z)\,dz = 2\pi j b_{-1} \qquad [193]$$

which agrees with Eq. 185.

The coefficient b_{-1} is called the *residue* of $f(z)$ in the point $z = z_0$. The value of a contour integral enclosing a singularity is, therefore, equal to $2\pi j$ times the residue of the function in this singularity. This result is referred to as Cauchy's *residue theorem.*

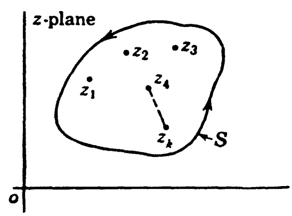


FIG. 14. The integral of a function about S enclosing several poles is determined by the sum of the residues.

When the contour S encloses more than one singularity, that is, if the function $f(z)$ is singular at several points $z_1, z_2, \cdots z_k$ within the enclosed region, as indicated in Fig. 14, the contour S may be replaced by k separate contours each enclosing one of the singularities, and the value of the contour integral around S is seen to be given by the sum of individual contour integrals around the k separate contours. It becomes clear that in this case

$$\oint_S f(z)\,dz = 2\pi j(b_{-1}^{(1)} + b_{-1}^{(2)} + \cdots + b_{-1}^{(k)}) \qquad [194]$$

in which $b_{-1}^{(1)}, b_{-1}^{(2)}, \cdots$ are the residues of $f(z)$ in the points $z_1, z_2, \cdots$ respectively.

The contraction of the contour S to the several separate contours about $z_1, z_2, \cdots$ may be visualized through supposing S to be a rubber band which is shrunk in the manner indicated in Fig. 15. The contributions coming to the net result from those portions of the shrunken contour which in the limit become superimposed and are traversed in opposite directions evidently cancel.

As a practical means for evaluating a given contour integral, of course, this method is useless unless some way is found for determining the residue, which is the coefficient of the first term in the principal part, Eq. 168, of the Laurent expansion. When several singularities are enclosed by the contour S, the residues of the function $f(z)$ are usually found separately for each of the singularities. If it is possible to find the first term in the principal part of the Laurent expansion for the immediate vicinity surrounding each singularity, this objective is accomplished.

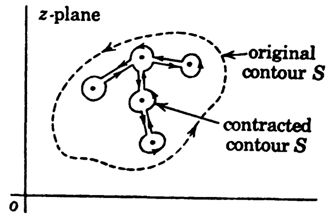


FIG. 15. The contribution of each pole to the integral is accounted for separately by shrinking the contour.

The following method for evaluating the residue is useful in many cases. If the singularity of $f(z)$ in z_0 is an ordinary pole, the reciprocal of this function

$$\phi(z) = \frac{1}{f(z)} \qquad [195]$$

is regular in z_0 and may there be expanded in the Taylor series

$$\phi(z)=\phi(z_0)+\phi'(z_0)\cdot(z-z_0)+\tfrac{1}{2}\phi''(z_0)\cdot(z-z_0)^2+\cdots \qquad [196]$$

in which the primes denote differentiation with respect to z.

The detailed process now varies according to the order of the pole of $f(z)$ in z_0. If this pole is of the first order, $\phi(z_0) = 0$, but $\phi'(z_0)$ is not zero. Division of the series 196 into unity (by the ordinary process of long division) yields

$$\frac{1}{\phi(z)} = f(z) = \frac{1}{\phi'}(z-z_0)^{-1} - \frac{1}{2}\frac{\phi''}{(\phi')^2}(z-z_0)^0 + \left\{\frac{1}{4}\frac{(\phi'')^2}{(\phi')^3} - \frac{1}{6}\frac{\phi'''}{(\phi')^2}\right\}(z-z_0) - \cdots \qquad [197]$$

This is the Laurent expansion of $f(z)$ about the point z_0. Hence the residue in this case is

$$b_{-1} = \left(\frac{d\phi}{dz}\right)^{-1}_{z=z_0} \qquad [198]$$

If the pole of $f(z)$ in z_0 is of the second order, $\phi(z_0) = 0$ and $\phi'(z_0) = 0$, but $\phi''(z_0) \neq 0$. Division of the series 196 into unity then reads

$$\frac{1}{\phi(z)} = f(z) = \frac{2}{\phi''}(z-z_0)^{-2} - \frac{2}{3}\frac{\phi'''}{(\phi'')^2}(z-z_0)^{-1} + \left\{\frac{2}{9}\frac{(\phi''')^2}{(\phi'')^3} - \frac{1}{6}\frac{\phi''''}{(\phi'')^2}\right\}(z-z_0)^0 - \cdots \qquad [199]$$

from which the residue is seen to be

$$b_{-1} = -\frac{2}{3}\frac{\left(\dfrac{d^3\phi}{dz^3}\right)_{z=z_0}}{\left(\dfrac{d^2\phi}{dz^2}\right)^2_{z=z_0}} \tag{200}$$

When the pole is of higher order, the expression for the residue becomes increasingly more complicated, but the method of evaluation remains the same.

The detailed aspects of the process of evaluating residues may, of course, be varied in a great many ways, and the most expeditious course depends entirely upon the form of the specific function in hand. Additional ways of approaching the problem may, in the course of the solution of specific examples, suggest still other variations.

If $f(z)$ has an sth order pole in $z = z_0$, then

$$\psi(z) = (z - z_0)^s f(z) \tag{201}$$

is regular in this point. Hence it possesses the Taylor expansion

$$\psi(z) = \psi(z_0) + \psi'(z_0)\cdot(z - z_0) + \tfrac{1}{2}\psi''(z_0)\cdot(z - z_0)^2 + \cdots \tag{202}$$

Since

$$f(z) = \frac{\psi(z)}{(z - z_0)^s} \tag{203}$$

substitution of the Taylor series 202 for $\psi(z)$ yields the Laurent expansion of $f(z)$ about z_0. It is then clear that the residue of $f(z)$ in z_0 is given by

$$b_{-1} = \frac{1}{(s-1)!}\left(\frac{d^{s-1}\psi}{dz^{s-1}}\right)_{z=z_0} \tag{204}$$

Sometimes the function $\psi(z)$ in Eq. 201 is more conveniently expressed as the product of two simpler functions:

$$\psi(z) = \xi(z)\cdot\eta(z) \tag{205}$$

Suppose that

$$\xi(z) = \alpha_0 + \alpha_1(z - z_0) + \alpha_2(z - z_0)^2 + \cdots \tag{206}$$

and

$$\eta(z) = \beta_0 + \beta_1(z - z_0) + \beta_2(z - z_0)^2 + \cdots \tag{207}$$

are the Taylor expansions for these functions about $z = z_0$. Since the residue of $f(z)$ in z_0 is the coefficient of the term containing $(z - z_0)^{s-1}$ in the product $\xi \cdot \eta$, it is seen that

$$b_{-1} = \alpha_0\beta_{s-1} + \alpha_1\beta_{s-2} + \cdots + \alpha_{s-1}\beta_0 \tag{208}$$

In numerous practical problems the function $f(z)$ in the integral 185

has a finite number of ordinary poles in the finite z-plane, all of which are enclosed by the boundary S. At the point infinity, the function may or may not be singular. If the complex plane is thought of as replaced by the surface of the complex sphere, it is possible to regard the contour S either as one enclosing all the singularities of $f(z)$ which occur for finite z-values or as one which encloses none of these singularities but merely surrounds the point at infinity. In other words, the region which ordinarily is regarded as being external to the contour S may alternatively be interpreted as being the enclosed region. The latter is, of course, traversed in the opposite sense, but for the moment this fact is of secondary importance. If it happens that the function $f(z)$ is regular at infinity, one is confronted with the peculiarity that the integral around a contour enclosing no singularities is nevertheless not equal to zero. The unique feature about this situation, however, is that the region in question contains the point at infinity. Hence one must conclude that the residue of a function in the point at infinity is not necessarily zero if the function is regular there. In fact, the residue of $f(z)$ at infinity is equal to the negative sum of its residues in all its singularities which occur for finite z-values.

A simple example may illustrate this point more specifically. Suppose $f(z) = 1/z$. This function has a simple pole at $z = 0$ (with the residue unity) and is regular everywhere else. At infinity the function has a simple zero. A circular path enclosing the origin may alternatively be regarded as a circular contour enclosing the point at infinity. If these paths are separately traversed in their counterclockwise directions, the values of the resulting integrals are $\pm 2\pi j$ respectively. Notwithstanding the fact that $f(z)$ is regular at infinity, it is seen that the contour integral enclosing this point has a nonzero value.

Conversely, one cannot conclude that the integral has a nonzero value for a contour enclosing the point at infinity if the function there has a simple pole. For example, if $f(z) = z$, the contour integral evidently has the value zero.

The residue of a function in the point at infinity cannot be evaluated by any of the processes which apply to finite points. One might suppose that such an evaluation could be accomplished through first introducing the change of variable indicated by the substitution $\zeta = 1/z$, which interchanges the origin with the point at infinity, and then proceeding in the normal fashion. Again the above example for the function $f(z) = 1/z$ shows that this method is obviously incorrect.

16. The partial fraction expansion of rational functions

If $f(z)$ is a rational function, then, as pointed out in Art. 13, it has a finite number of poles in the entire z-plane. This number, the pole at

infinity being excluded if present, may be denoted by n, and the z-values corresponding to the poles of $f(z)$ by $z_1, z_2, \cdots z_n$. The principal parts of Laurent expansions for the function about the points $z_1, z_2, \cdots z_n$ are denoted by $h_1, h_2, \cdots h_n$ respectively. The function

$$f(z) - h_1(z) - h_2(z) - \cdots - h_n(z) = g(z) \qquad [209]$$

must be an entire rational function, that is, a polynomial in z whose highest power equals the order of the pole of $f(z)$ at infinity. This fact is recognized through observing, for example, that $f(z) - h_1(z)$ must be regular at the point $z = z_1$ because its Laurent expansion about z_1 has no principal part and hence is a Taylor expansion. However, the function $f(z) - h_1(z)$ still has poles at the points $z_2, z_3, \cdots z_n$. Next, the function $f(z) - h_1(z) - h_2(z)$ is seen to be regular at the points z_1 and z_2, and therefore has only the poles at $z_3, z_4, \cdots z_n$, and so forth.

Transposing the principal parts in Eq. 209 to the right-hand side, one obtains the representation

$$f(z) = h_1(z) + h_2(z) + \cdots + h_n(z) + g(z) \qquad [210]$$

in which each term places one of the poles of $f(z)$ in evidence. It is an explicit representation of the function in the form of a linear superposition of the individual contributions of its singularities. This representation for the function $f(z)$ is known as its *partial fraction expansion*.

More specifically, if

$$f(z) = \frac{p(z)}{q(z)} = \frac{\alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_m z^m}{\beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_n z^n} \qquad [211]$$

and it is assumed that all the roots of $q(z)$ are distinct, all the poles of $f(z)$ are simple. These roots may be denoted by $z_1, z_2, \cdots z_n$. The principal parts of $f(z)$ in its poles then have the form

$$h_\nu(z) = \frac{b_{-1}{}^{(\nu)}}{z - z_\nu} \qquad [212]$$

in which $b_{-1}{}^{(1)}, b_{-1}{}^{(2)}, \cdots b_{-1}{}^{(n)}$ are the corresponding residues.

Applying Eq. 198 of the previous article, and noting that according to Eq. 195 $\phi(z)$ is in this example $\phi = q/p$, one finds

$$b_{-1}{}^{(\nu)} = \left\{\frac{p(z)}{\dfrac{dq}{dz}}\right\}_{z=z_\nu} \qquad [213]$$

The derivative of $q(z)$ for $z = z_\nu$ may be further evaluated through noting that the factored form of this polynomial reads

$$q(z) = \beta_n(z - z_1)(z - z_2) \cdots (z - z_n) \qquad [214]$$

and hence that

$$\left(\frac{dq}{dz}\right)_{z=z_\nu} = \beta_n(z_\nu - z_1)(z_\nu - z_2)\cdots(z_\nu - z_{\nu-1})(z_\nu - z_{\nu+1})\cdots(z_\nu - z_n) \tag{215}$$

which may alternatively be written

$$\left(\frac{dq}{dz}\right)_{z=z_\nu} = \left\{\frac{q(z)}{z - z_\nu}\right\}_{z=z_\nu} \tag{216}$$

The way in which Eq. 215 is arrived at may readily be seen through first regarding $q(z)$ in Eq. 214 as being in the form

$$q(z) = (z - z_\nu)\cdot q^*(z) \tag{217}$$

in which $q^*(z)$ is the right-hand side of Eq. 214 with the factor $(z - z_\nu)$ missing. Now applying the rule for the derivative of a product to $q(z)$ as represented in Eq. 217, one finds that

$$\frac{dq}{dz} = (z - z_\nu)\frac{dq^*}{dz} + q^*(z) \tag{218}$$

and hence that

$$\left(\frac{dq}{dz}\right)_{z=z_\nu} = q^*(z_\nu) \tag{219}$$

which agrees with Eqs. 215 and 216.

In view of these considerations it becomes clear that the residues as given by Eq. 213 may alternatively be written in the form

$$b_{-1}^{(\nu)} = [(z - z_\nu)\cdot f(z)]_{z=z_\nu} \tag{220}$$

These results are unchanged if the polynomial $q(z)$ has a zero root, that is, if $\beta_0 = 0$. They are restricted to the case of simple poles, of course, since the roots of $q(z)$ are, in the above analysis, assumed to be distinct. When multiple roots occur, methods similar to those discussed in the previous article for evaluating residues at multiple-order poles must be used to determine the principal parts $h_\nu(z)$ in Eq. 210.

If all the poles of $f(z)$ are assumed to be simple, the results stated by Eqs. 212 and 220 determine the partial fraction expansion 210 except for the rational integral function $g(z)$. This function is found from the form of $f(z)$ given by Eq. 211 in the following way. If $m \leqq n$, then $f(z)$ is regular at infinity, and the entire rational function $g(z)$ reduces to a constant. More specifically, if $m = n$, then

$$g(z) = \frac{\alpha_n}{\beta_n} \tag{221}$$

whereas if $m < n$, then $g(z)$ is identically zero.

On the other hand if $m > n$ and $m - n = s$, then $f(z)$ has an sth order pole at infinity. The function $g(z)$ is found through dividing $q(z)$ into $p(z)$ by long division so as to get

$$\frac{p(z)}{q(z)} = \gamma_s z^s + \gamma_{s-1} z^{s-1} + \cdots + \gamma_1 z + \gamma_0 + \frac{p^*(z)}{q(z)} \qquad [222]$$

in which the remainder polynomial $p^*(z)$ has the degree $n - 1$. Then

$$g(z) = \gamma_0 + \gamma_1 z + \cdots + \gamma_s z^s \qquad [223]$$

For $m = n$ this process yields $g(z) = \gamma_0 = \alpha_n / \beta_n$, whereas for $m < n$ it is clear that $g(z) \equiv 0$ as stated above.

These considerations amount to putting $f(z)$ into the form

$$f(z) = g(z) + \frac{p^*(z)}{q(z)} = g(z) + f^*(z) \qquad [224]$$

in which $f^*(z)$ has a simple zero at infinity but contains the same poles as $f(z)$ for the rest of the z-plane. In these poles $f^*(z)$ has the same principal parts $h_\nu(z)$ as the function $f(z)$, and these principal parts are found according to Eqs. 212, 213, and 220 by use of either $p(z)$ and $f(z)$ or $p^*(z)$ and $f^*(z)$ in these expressions, whichever appear to be more expedient.

17. Multivalued functions; branch points and Riemann surfaces

A multivalued function with which the reader undoubtedly has some acquaintance is the logarithm.* This function is defined by the integral

$$\ln z = \int_1^z \frac{d\zeta}{\zeta} \qquad [225]$$

in which the path of integration in the ζ-plane extends from the point $\zeta = 1$ to the point $\zeta = z$, but otherwise remains arbitrary. The integrand is the function

$$f(\zeta) = \frac{1}{\zeta} \qquad [226]$$

which is regular everywhere except at the origin ($\zeta = 0$) where it has a simple pole with the residue unity.

Hence the value of the integral is not affected by a deformation of the path of integration so long as no portion of this path is allowed to sweep

*Unless otherwise specified it is understood that the natural logarithm is implied. This is denoted by ln z to distinguish it from the Briggs logarithm, which is written log z.

across the origin. With reference to Fig. 16, one has

$$\int_{P_1} \frac{d\zeta}{\zeta} - \int_{P_2} \frac{d\zeta}{\zeta} = \oint_1 \frac{d\zeta}{\zeta} \qquad [227]$$

in which the subscript 1 on the closed contour integral indicates that the origin is encircled once in the counterclockwise direction. From Art. 15 it is seen that therefore

$$\int_{P_1} \frac{d\zeta}{\zeta} - \int_{P_2} \frac{d\zeta}{\zeta} = 2\pi j \qquad [228]$$

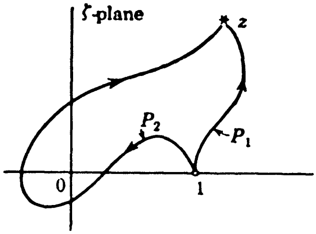


FIG. 16. The integral on closed path P_1-P_2 yields $2\pi j$ since the branch point at the origin is encircled once in counterclockwise sense.

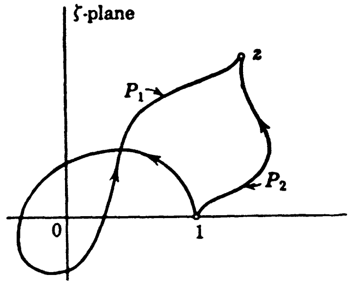


FIG. 17. The integral on closed path P_1-P_2 again yields $2\pi j$. Path is different from that in Fig. 16 but sense around branch point is the same.

The same result is true for the two alternative paths shown in Fig. 17. In general, if the path P_1 encircles the origin in the counterclockwise direction n more times than the path P_2 does,

$$\int_{P_1} \frac{d\zeta}{\zeta} - \int_{P_2} \frac{d\zeta}{\zeta} = 2\pi n j \qquad [229]$$

Hence if in conjunction with the definition 225 no statement is made relative to the path of integration, the value of the logarithm is determined only within an arbitrary integer number of $2\pi j$'s. Its multivaluedness is thus apparent.

This result is also readily obtained from the more familiar definition of the logarithm, according to which

$$z = e^{\ln z} \qquad [230]$$

If $\ln z$ is here replaced by $\ln z + 2\pi n j$, the value of the exponential differs only by the factor $e^{2\pi n j}$, which has the value unity. The view stated in the preceding paragraph, however, gives some pictorial signifi-

cance to this multivaluedness, whereas the more elementary reasoning precludes the possibility of such an interpretation.

For further elucidation, the following more detailed representation is helpful. The complex number z is written in the polar form

$$z = re^{j\phi} \tag{231}$$

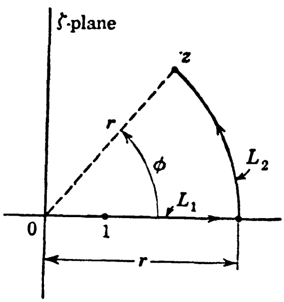


FIG. 18. Integral of $1/z$ along L_1 contributes to real part of logarithm: that along L_2 contributes to imaginary part.

As shown in Fig. 18, the path of integration is assumed to consist of the portion L_1 coincident with the real axis from 1 to r, followed by the portion L_2, which is an arc drawn from the point r on the real axis to z.

The variable of integration is also written in the polar form

$$\zeta = \rho e^{j\theta} \tag{232}$$

whence

$$d\zeta = \frac{\partial\zeta}{\partial\rho}d\rho + \frac{\partial\zeta}{\partial\theta}d\theta \tag{233}$$

or

$$d\zeta = e^{j\theta}\,d\rho + j\rho e^{j\theta}\,d\theta = (d\rho + j\rho\,d\theta)e^{j\theta} \tag{234}$$

Then

$$\frac{d\zeta}{\zeta} = \frac{d\rho}{\rho} + j\,d\theta \tag{235}$$

The integral 225, separated into two parts corresponding respectively to the portions L_1 and L_2 of the resultant path of integration, reads

$$\ln z = \int_{\substack{1\\L_1}}^{r}\frac{d\rho}{\rho} + j\int_{\substack{0\\L_2}}^{\phi}d\theta \tag{236}$$

Now

$$\int_1^r \frac{d\rho}{\rho} = \ln r \tag{237}$$

is the logarithm of the magnitude of z, whereas

$$\int_0^\phi d\theta = \phi \tag{238}$$

is the angle of z. Hence

$$\ln z = \ln r + j\phi \tag{239}$$

The multivaluedness of the logarithm is seen to result from the addition or subtraction of an integer number of complete revolutions to the path

L_2. This operation simply adds positive or negative integer multiples of 2π to the upper limit of the integral 238 and hence to ϕ appearing in Eq. 239. Written in more complete form, this relation, therefore, reads

$$\ln z = \ln r + j\phi + j2\pi n \qquad [240]$$

in which n is any positive or negative integer.

For the detailed discussion of the multivalued character of the logarithm it is expedient to define the value of Eq. 240 for $n = 0$, namely,

$$\mathrm{Ln}\, z = \ln r + j\phi \qquad [241]$$

as the *principal value* of $\ln z$, and write

$$\ln z = \mathrm{Ln}\, z + j2\pi n \qquad [242]$$

Ordinarily in speaking of the logarithm one has in mind the principal value only.

Now that the multivalued character of the logarithm and the reasons for it are established, attention may be given to the question of how the ambiguity can be taken care of when this function enters into some problem, for example, in the consideration of the mapping of $w = \ln z$ in the w- and z-planes. To a given point in the z-plane there correspond an infinite number of points in the w-plane, all of which have the same real part whereas their imaginary parts differ by multiples of 2π.

Such a set of points in the w-plane is presented in Fig. 19, which shows that the entire z-plane is mapped in any one of the oppositely cross-hatched strips, 2π units wide. Horizontal lines in the w-plane correspond to radial lines (ϕ = constant) in the z-plane; vertical lines in the w-plane correspond to concentric circles about the origin (r = constant) in the z-plane. The system of concentric circles and radial lines in the z-plane is, therefore, transformed into a rectangular grid in the w-plane.

The fact that the locus joining a set of points in the w-plane corresponding to the same z-value (like the vertical dotted line in Fig. 19) becomes a circular locus in the z-plane which winds around and around the origin suggests that the multivaluedness may be eliminated artificially through conceiving the z-plane in the form of a winding surface comprising an infinite number of superimposed leaves which have a common central point at the origin and simulate a winding staircase of infinite width in which the steps are replaced by a smooth continuous ramp. As the slope of this winding surface is made smaller and smaller, the spacings between the leaves (or successive elevations of the ramp) ultimately become negligibly small.

The one original z-value representing an infinite number of w-values differing by $j2\pi n$ is now separated into a uniquely corresponding infinite number of z-values which lie directly above and below each other on the

various stages of the ramp or levels of the continuous winding surface. Each level, or leaf, of the surface carries one of the separate z-values which now corresponds uniquely to one of the w-values. To go from the z-value representing a given w-value to that representing the value $w \pm j2\pi$, one must follow the ramp once around the origin in the counterclockwise or clockwise direction respectively. The winding sense of the ramp is such that the increment in w is $+j2\pi$ for one revolution in the counterclockwise direction.

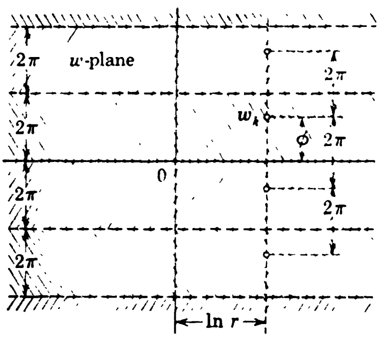


FIG. 19. Sections of the w-plane each of which corresponds to a leaf of the Riemann surface of $\ln z$ in the z-plane.

This hypothetical surface which thus effectively renders the function single-valued is a *Riemann surface*. The origin or common point about which the surface winds is called the *winding point* or *branch point* of the Riemann surface. It evidently is a singularity, for the logarithm function does not have a finite value there.* With the exception of this point and the one at infinity, the function $w = \ln z$ is regular in the entire Riemann surface, and its integral around any closed contour is zero because it is impossible to wind about the origin in the same sense a nonzero number of times and return to a given starting point.

The logarithm is an illustration of a multivalued transcendental function. The Riemann surface for such a function has an infinite number of leaves because the multivaluedness is infinite. The inverse trigonometric functions $\tan^{-1} z$, $\sin^{-1} z$, $\cos^{-1} z$, etc., are other examples.

A class of multivalued functions having a finite degree of multivaluedness, and hence possessing Riemann surfaces with a finite number of leaves, is found in the *algebraic* functions.† A simple example is the function

$$w = \sqrt{z - z_0} \tag{243}$$

In view of the preceding discussion it is convenient to write this expression in the form

$$w = e^{1/2\,\mathrm{Ln}\,(z-z_0)+j\pi n} \tag{244}$$

*It should not be inferred that branch points necessarily are singularities of this kind. It is possible (as in the case of many algebraic functions) for the function to have a finite and unique value at a branch point.

†Not all algebraic functions are necessarily multivalued. The rational functions, for example, are single-valued algebraic functions.

For a given value of z, this function has two values, one for $n = 0$ and the other for $n = 1$. For $n = 2$ the same value again obtains as for $n = 0$, and for $n = 3$ the $(n = 1)$-value is repeated, etc. The Riemann surface for this function, therefore, has but two leaves. The branch point is located at $z = z_0$.

The structure of the Riemann surface about $z = z_0$ is similar to that of the logarithm function about its branch point, but the complete surface must now be so constructed that the same leaf is regained after two complete revolutions about the branch point. The first step toward clarifying this picture involves a recognition of the fact that the point at infinity has the same character as the point z_0. This circumstance is somewhat easier to see if the variable is changed by the transformation

$$z' = \frac{1}{z - z_0} \qquad [245]$$

Then Eqs. 243 and 244 become

$$w = (z')^{-1/2} = e^{-1/2\,\mathrm{Ln}\, z' - j\pi n} \qquad [246]$$

in which the point $z' = 0$ corresponds to $z = \infty$.

Since the point at infinity must be included in the present visualization process, it may be a little easier if the z-plane is replaced by its associated complex sphere. This sphere is imagined to have a double surface. Both surfaces are slit from the point $z = z_0$ to the point at infinity along a path which is arbitrary except that it shall have no crossover points. Throughout the entire length of this cut, the top leaves of the double surface are now imagined to be joined to the bottom leaves on the opposite sides of the cut (a physical impossibility, of course, but not beyond the powers of a good imagination) so that crossing the cut in either direction effects a transfer from the upper to the lower leaf of the Riemann surface or vice versa. This continuous duplex ramp is called a *branch cut*.

The resulting duplex surface evidently has the properties required for the unique mapping of the function given by Eq. 243. The two z-values corresponding to the two w-values that differ by a factor $e^{j\pi}$ lie above one another on the two leaves of the Riemann surface, and to go from one of these z-values to the other it is necessary to cross the branch cut and pass completely around either the branch point at $z = z_0$ or that at $z = \infty$. The same z-value may be regained by passage around a branch point an even number of times.

The function

$$w = \sqrt{(z - z_1)(z - z_2) \cdots (z - z_k)} \qquad [247]$$

is also double-valued, and its Riemann surface again has two leaves. The branch points are $z_1, z_2, \cdots z_k$. For very large values of z the function

behaves like $z^{k/2}$, and hence it is clear that the point at infinity is also a branch point if k is odd.

The branch cuts of the Riemann surface are made along paths joining the points z_1 and z_2, z_3 and z_4, etc. These paths cannot have any crossover points. In particular, for $k = 2$ the branch points are $z = z_1$ and $z = z_2$, which are joined by a branch cut in a manner exactly analogous to that discussed for the points $z = z_0$ and $z = \infty$ in the preceding example. In fact, the two-leaved Riemann surface for the function 247 with $k = 2$ is in all respects similar to that for the function 243 except that the second branch point lies at a finite z-value. A closed path which encircles both branch points z_1 and z_2 remains on the same leaf of the Riemann surface; that is, a given point z on this closed path is regained after a single traversal.

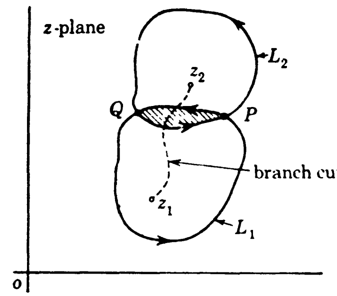


FIG. 20. A point traversing L_1 and L_2 in succession must return to the same leaf of the Riemann surface.

This situation may be clarified by reference to Fig. 20. A traversal enclosing both points z_1 and z_2 which begins and ends at P must be equivalent to a traversal around the closed path L_1 followed by a traversal around the closed path L_2, because the difference amounts only to the traversal from P to Q and back to P, around the shaded area which contains no branch point. Inasmuch as a separate traversal around L_1 or around L_2 effects a transition from one leaf of the Riemann surface to the other, it is clear that a given point traversing these two circuits in succession must return to the same leaf of the Riemann surface.

Analogous reasoning shows that the cases $k = 3$ and $k = 4$ have similar Riemann surfaces. For $k = 4$ the finite point z_4 replaces the branch point which for $k = 3$ occurs at infinity. The surface has two branch cuts, and a closed circuit which surrounds any two branch points returns to a given point after one traversal. The extension of this reasoning to the interpretation of the two-leaved Riemann surfaces for larger values of k is straightforward.

An extension of the same reasoning likewise leads to the required structure of the Riemann surface for the function

$$w = \sqrt[m]{z - z_0} \tag{248}$$

or

$$w = e^{\frac{1}{m}\operatorname{Ln}(z - z_0) + j\frac{2\pi n}{m}} \tag{249}$$

This function is m-valued, has a branch point of the $(m - 1)$th order

at the point $z = z_0$, and possesses a Riemann surface having m leaves. The point at infinity again has the same character as the point $z = z_0$, and these two points are joined by a slightly more complicated branch cut. At this branch cut the top leaf of the Riemann surface is joined to the one located below it on the opposite side of the cut, whereas the second leaf on the original side is joined to the third leaf from the top on the opposite side, etc. Finally, the bottom-most leaf on the original side of the cut is joined to the top leaf on the opposite side. The result is that a path must encircle either branch point m times before a given point on one of the leaves can be regained.

It may be useful to observe that the function $w = \sqrt[m]{z}$ has a Riemann surface entirely similar in its structure to that of the function $w = \ln z$ except that the latter has an infinite number of leaves. The branch cut in the case of the function $w = \sqrt[m]{z}$ is a necessary concept only because some mechanism must be imagined whereby the bottom leaf rejoins the top leaf of the surface. Actually leaving this mechanism entirely to the imagination is more effective than trying to formulate some sort of piecing and pasting process between the leaves which must afterward be apologized for because of the mechanical impossibility of carrying out such a scheme in a physical model.

A mathematician prefers not to be annoyed with the physical difficulties involved in the visualization of a branch cut, the more so since it implies the existence of a definite path along which the passage from one leaf to another takes place. According to the true conception of a Riemann surface, such passage is not to be regarded as localized in a branch cut. Rather, one is to regard the invention of a branch cut as made necessary only by reason of the inadequacy of one's habitual conception of space to comprehend the mechanism of the Riemann surface.

In the function $w = \ln z$ this difficulty does not appear, but it is replaced by the equally difficult conception of an infinite number of leaves in the Riemann surface. In terms of the complex sphere associated with the z-plane, the south pole (origin) and the north pole (infinity) are branch points of infinite order. The Riemann surface is a continuous succession of spherical shells winding round and round a polar axis, somewhat on the order of a snail's shell only that the winding pitch is zero and the number of revolutions infinite. The surface for the function $w = \sqrt[m]{z}$ has the same structure but comprises only m revolutions, after which identity with the original leaf is again established.

18. Algebraic functions; more about the classification of functions

From the discussion just given it is clear that if the function $w = f(z)$ has a branch point of the $(m - 1)$th order in $z = z_0$, then, by use of the

substitution

$$z - z_0 = re^{j\phi} \tag{250}$$

one may study the function

$$f(z) = f(z_0 + re^{j\phi}) \tag{251}$$

for an appropriate value of r on the various leaves of its Riemann surface through letting ϕ vary continuously from some initial value ϕ_0 to a final value $\phi_0 + 2m\pi$. A further increase in ϕ will merely yield the same values over again.

The m leaves of the Riemann surface for this vicinity of the point z_0 are encountered in cyclic order in the intervals

$$\phi_0 + 2(\nu - 1)\pi < \phi < \phi_0 + 2\nu\pi \tag{252}$$

with $\nu = 1, 2, \cdots m$.

The m-values of the multivalued function $f(z)$ may correspondingly be denoted by

$$w_\nu = f(z_0 + re^{j[\phi + 2(\nu - 1)\pi]}) \tag{253}$$

with $\phi_0 < \phi < \phi_0 + 2\pi$.

The quantities $w_1, w_2, \cdots w_m$, which represent the function on the various leaves of its Riemann surface, are called the *branches* of the function $f(z)$. They evidently form a cyclic group, since, for the interval $\phi_0 + 2\pi < \phi < \phi_0 + 4\pi$, for example, w_1 replaces w_2, w_2 replaces w_3, and so forth, and w_m replaces w_1.

In terms of these branches of the function $w = f(z)$, an important property of the branch point of finite order may now be stated, namely, that the same limiting value of the function results for the limit $z \to z_0$ regardless of which one of the branches $w_1, w_2, \cdots$ etc., is chosen in the process of evaluating this limit. In other words, the value of $f(z)$ in the branch point may be approached from any one of the m leaves of the Riemann surface. Since the branch point thus yields a unique value for the function, it may be included as a point of this surface.

It is possible for a multivalued function to have this character in the vicinity of a point $z = z_0$ and there possess an infinite or finite value, either of which is definitely determinable. A singularity of this sort has a more general character than that of either a branch point or an ordinary pole, since it embraces both these as special cases. It is called an *algebraic singularity* inasmuch as it represents the only kind of singularity found in that class of functions known as the *algebraic functions*.

In order to study the behavior of an m-valued function in the vicinity of an algebraic singularity at the point $z = z_0$, one may make the change of variable indicated by

$$t = (z - z_0)^{1/m} \tag{254}$$

Then

$$w = f(z) = \psi(t) \qquad [255]$$

becomes a single-valued function of the complex variable t, since the latter has taken over the m-valued character in terms of the original variable z. If the function $\psi(t)$ possesses an nth order pole at the point $t = 0$ (corresponding to $z = z_0$) then, for the vicinity of this point, it may be represented by a Laurent expansion whose descending part contains n terms. Hence it becomes clear that the function $f(z)$ admits the following expansion in the vicinity of an algebraic singularity:

$$f(z) = \sum_{\nu=-n}^{\infty} C_\nu (z - z_0)^{\nu/m} \qquad [256]$$

This singularity is an ordinary pole of the nth order for the function $f(z)$ if $m = 1$. It represents an ordinary branch point of the $(m - 1)$th order if $n = 0$.

In general, that is for $m > 1$ and $n > 0$, the function $f(z)$ is said to possess a singularity at the point $z = z_0$ which is simultaneously a pole of the order n and a branch point of the order $m - 1$. For example, the function $f(z) = \sqrt{z}$ has a branch point of order 1 at $z = 0$, whereas at $z = \infty$ it has simultaneously a simple pole and a branch point of order 1. The same statement, with an interchange of reference to the points $z = 0$ and $z = \infty$, applies to the function $f(z) = 1/\sqrt{z}$. If n is infinite, the function has an essential singularity at the branch point of order $m - 1$. On the other hand, if m is infinite, the function is said to have a *logarithmic branch point* (it is then no longer an algebraic function, for the latter can have branch points of finite order only).

Algebraic functions are defined as functions possessing only a *finite* number of algebraic singularities. If the function is m-valued, it is not necessary that all or that even one of its singularities also be a branch point of the order $(m - 1)$, but a sufficient number of these singularities must be branch points of such order and distribution as will insure that the m leaves of the associated Riemann surface form a connected system. For any nonsingular z-value the function $f(z)$ possesses the m-values indicated by

$$f(z) = w_1, w_2, \cdots w_m \qquad [257]$$

which represent the function on the correspondingly numbered leaves of its Riemann surface. In terms of these branches of the function $f(z)$ one may form the system of symmetrical functions

$$\begin{aligned} \psi_1(z) &= w_1 + w_2 + \cdots + w_m \\ \psi_2(z) &= w_1w_2 + w_1w_3 + \cdots + w_{m-1}w_m \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \psi_m(z) &= w_1w_2 \cdots w_m \end{aligned} \qquad [258]$$

In view of the fact that the branches $w_1, w_2, \cdots w_m$ form a cyclic group, it follows that all the symmetrical functions $\psi_1, \psi_2, \cdots \psi_m$ are single-valued. That is, as the point z is allowed to move along any path on the Riemann surface, avoiding only the singularities of $f(z)$, the functions $w_1, w_2, \cdots w_m$ can merely interchange their identities, and since the functions $\psi_1, \psi_2, \cdots \psi_m$ are given by symmetrical combinations of the elements $w_1, w_2, \cdots w_m$, they are not affected by such interchanges. Furthermore, they can have only algebraic singularities because this is the only kind that the branches $w_1, w_2, \cdots w_m$ possess, and the ψ's are formed from the w's by the processes of addition and multiplication alone. Algebraic singularities in the case of single-valued functions, however, must be ordinary poles. Hence the functions $\psi_1(z), \psi_2(z), \cdots \psi_m(z)$ must be rational.

Now it is to be recalled from the theory of algebraic equations that the symmetrical functions defined by Eq. 258 satisfy the equation

$$\begin{aligned}(w - w_1)(w - w_2) \cdots (w - w_m) \\ = w^m - \psi_1(z) \cdot w^{m-1} + \cdots + (-1)^m \psi_m(z) = 0 \end{aligned} \quad [259]$$

It is also recognized that the functions 258 have a common denominator equal to the product of the denominators of the branches $w_1, w_2, \cdots w_m$, and that this common denominator must be a rational entire function of z, from the last of Eqs. 258, or alternately because the branches $w_1, w_2, \cdots w_m$ individually are single-valued on their respective leaves of the Riemann surface and there possess singularities which, therefore, can be none other than ordinary poles. Multiplying Eq. 259 by this rational entire function, which is a finite polynomial $p_0(z)$, one finds

$$F(z,w) = p_0(z) \cdot w^m + p_1(z) \cdot w^{m-1} + \cdots + p_m(z) = 0 \quad [260]$$

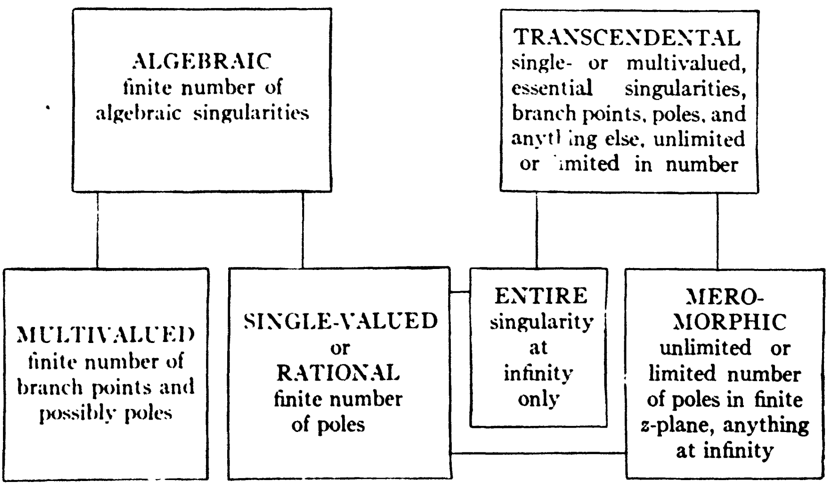
in which $p_0, p_1, \cdots p_m$ are finite polynomials in z.

The algebraic functions $w = f(z)$ are thus seen to be defined as the roots of an algebraic equation whose coefficients are ordinary polynomials in z. More precisely, the roots are the branches which collectively define a single algebraic function on the various leaves of its Riemann surface. The function $F(z,w)$ may never be reducible to the product of two or more factors having the same form as $F(z,w)$, since the vanishing of any factor alone would then satisfy Eq. 260, and hence several independent functions rather than a single one would be defined by this equation.

The branch points occur for those values of z for which Eq. 260 has coincident w-roots. The coincidence of roots requires that the discriminant be zero; and since this discriminant is a rational function of the polynomials $p_0, p_1, \cdots p_m$ and hence a rational function of z, it cannot vanish identically, but can do so only for a finite number of z-values. Hence the number of branch points of the algebraic function defined by Eq. 260 is finite.

The engineering student will find it helpful to visualize the Riemann surface as a set of m parallel metal sheets and the branch points as spot welds which join two or more of the sheets at isolated points. The distribution of these spot welds and the number of sheets held together by each must evidently be such that the m sheets are connected; otherwise instead of a single m-valued function, several functions of lesser order than m are defined.

The classification of functions in terms of the nature and distribution of their singularities, which is partially discussed in Art. 13, may now be viewed in a more thorough fashion. All analytic functions may be divided into two main classes, which are the algebraic and the transcendental. In other words, any function which is not algebraic belongs to the transcendental group, comprising single- or multivalued functions having essential singularities. Algebraic functions, on the other hand, may be further subdivided into single-valued and multivalued functions. The single-valued ones are identified with the rational functions, which include some of the entire functions, namely, the finite polynomials. The meromorphic functions, which are single-valued and may have any kind of singularity at infinity but must have only poles in the finite z-plane, include some of the functions in the transcendental group and all the functions in the rational group. The entire functions may likewise be regarded as a subclass in the transcendental group, but in addition they lay claim to some of the functions in the rational group. The following block diagram is intended to unify these remarks.



It is now possible to discuss somewhat more adequately the inversion of functions (which is introduced in Art. 3) with particular reference to the vicinity of saddle points described in Art. 14. Suppose a given single

valued function $w = f(z)$ has, at a point $z = z_0$, vanishing first and higher derivatives up to but not including the nth. For the vicinity of this point it then possesses the Taylor expansion

$$w = w_0 + a_n(z - z_0)^n + a_{n+1}(z - z_0)^{n+1} + \cdots \qquad [261]$$

with $a_n \neq 0$. Writing this equation in the form

$$\frac{w - w_0}{a_n} = (z - z_0)^n + \frac{a_{n+1}}{a_n}(z - z_0)^{n+1} + \cdots \qquad [262]$$

extracting the nth root, and introducing for abbreviation the variable

$$\tau = \left(\frac{w - w_0}{a_n}\right)^{1/n} \qquad [263]$$

one may put the series 262 into the form

$$\tau = (z - z_0)\left\{1 + \frac{a_{n+1}}{a_n}(z - z_0) + \frac{a_{n+2}}{a_n}(z - z_0)^2 + \cdots\right\}^{1/n} \qquad [264]$$

The bracketed expression, inclusive of the exponent $1/n$, may for the vicinity of the point $z = z_0$ (which is not a branch point) be expanded in a Taylor series having the form

$$1 + b_1(z - z_0) + b_2(z - z_0)^2 + \cdots \qquad [265]$$

This Taylor series represents the bracket function in Eq. 264 on only one of the leaves of its Riemann surface. The corresponding representations on the remaining $n - 1$ leaves of this surface, however, differ from 265 only by the factor

$$e^{j(2\nu\pi/n)} \text{ for } \nu = 1, 2, \cdots n - 1 \qquad [266]$$

Substituting 265 for the bracket expression in Eq. 264 yields

$$\tau = (z - z_0) + b_1(z - z_0)^2 + b_2(z - z_0)^3 + \cdots \qquad [267]$$

which, on the particular leaf of the Riemann surface in question, is a unique representation for $\tau(z)$ in the vicinity of the point $z = z_0$. Hence, according to the discussion in Art. 3, it possesses an inverse function whose Taylor series for the vicinity of $\tau = 0$ (which is a regular point) has the form

$$(z - z_0) = \beta_1\tau + \beta_2\tau^2 + \beta_3\tau^3 + \cdots \qquad [268]$$

The coefficients in this series for the inverse function may, for example, be found through first substituting the series for $(z - z_0)$ into Eq. 267 and, after arrangement in ascending powers of τ, obtaining

$$\tau = \beta_1\tau + (b_1\beta_1^2 + \beta_2)\tau^2 + (2b_1\beta_1\beta_2 + b_2\beta_1 + \beta_3)\tau^3 + \cdots \qquad [269]$$

whence, equating coefficients of like powers of τ, one has

$$\begin{aligned} \beta_1 &= 1, \\ \beta_2 &= -b_1, \\ \beta_3 &= 2b_1^2 - b_2 \\ &\cdots\cdots\cdots\cdots \end{aligned} \tag{270}$$

Substitution of the expression 263 for τ into Eq. 268 then gives

$$z = z_0 + \left(\frac{w - w_0}{a_n}\right)^{1/n} + \beta_2\left(\frac{w - w_0}{a_n}\right)^{2/n} + \beta_3\left(\frac{w - w_0}{a_n}\right)^{3/n} + \cdots \tag{271}$$

which is a representation for the inverse function $z = \phi(w)$ for the vicinity of the point $w = w_0$, corresponding to the point $z = z_0$. Equation 271 may be written

$$z = \phi(w_0) + c_1(w - w_0)^{1/n} + c_2(w - w_0)^{2/n} + \cdots \tag{272}$$

This, however, is recognized (according to the discussion leading to Eq. 256) as the expansion of an n-valued function in the vicinity of a branch point of the $(n-1)$th order at the point $w = w_0$. Hence it is established that if a given function $w = f(z)$ has a saddle point of the $(n-1)$th order at a point $z = z_0$, the inverse function $z = \phi(w)$ has a branch point of the $(n-1)$th order in the corresponding point $w = w_0$. If the point $z = z_0$ is a zero of the nth order for the function $w = f(z)$, the above analysis remains unaltered except that w_0 becomes zero. The inverse function $z = \phi(w)$ then has a branch point of the $(n-1)$th order at the *origin* in the w-plane instead of at the point $w = w_0$.

19. A THEOREM REGARDING THE NUMBER OF ZEROS AND POLES WITHIN A GIVEN REGION; THE FUNDAMENTAL LAW OF ALGEBRA

A given function $w = f(z)$ is assumed to have a zero of the order α in the point $z = z_0$. The Taylor expansion about this point then reads

$$f(z) = a_\alpha(z - z_0)^\alpha + a_{\alpha+1}(z - z_0)^{\alpha+1} + \cdots \tag{273}$$

and the derivative of $f(z)$ is given by

$$f'(z) = \alpha a_\alpha(z - z_0)^{\alpha-1} + (\alpha + 1)a_{\alpha+1}(z - z_0)^\alpha + \cdots \tag{274}$$

Dividing the series 273 into the series 274 by long division gives

$$\frac{f'(z)}{f(z)} = \alpha(z - z_0)^{-1} + c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + \cdots \tag{275}$$

This function, therefore, is seen to have a pole of the first order in the point $z = z_0$ with the residue α.

Alternatively, let it be supposed that the given function $f(z)$ has a pole

in z^* of the order β. Then its Laurent expansion about this point is given by

$$f(z) = b_{-\beta}(z - z^*)^{-\beta} + b_{-\beta+1}(z - z^*)^{-\beta+1} + \cdots \qquad [276]$$

and the derivative reads

$$f'(z) = -\beta b_{-\beta}(z - z^*)^{-\beta-1} - (\beta - 1)b_{-\beta+1}(z - z^*)^{-\beta} + \cdots \qquad [277]$$

By long division it is then found that

$$\frac{f'(z)}{f(z)} = -\beta(z - z^*)^{-1} + d_0 + d_1(z - z^*) + d_2(z - z^*)^2 + \cdots \qquad [278]$$

from which this function is seen to have a simple pole in the point $z = z^*$ with the residue $-\beta$.

According to the residue theorem it follows from these considerations that

$$\oint_C \frac{f'(z)}{f(z)}\, dz = 2\pi j(\alpha - \beta) \qquad [279]$$

where the path C is assumed to enclose no other zeros or poles of the function $f(z)$ except those at the points z_0 and z^*.

In general it is seen that if the path C encloses zeros of $f(z)$ having the orders $\alpha_1, \alpha_2, \cdots \alpha_k$, and poles having the orders $\beta_1, \beta_2, \cdots \beta_s$, and if

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k = N \qquad [280]$$

and

$$\beta_1 + \beta_2 + \cdots + \beta_s = P \qquad [281]$$

then

$$\frac{1}{2\pi j}\oint_C \frac{f'(z)}{f(z)}\, dz = N - P \qquad [282]$$

The integer N is equal to the total number of enclosed zeros of the function $f(z)$, each one being counted as often as its order requires; the integer P is equal to the total number of enclosed poles of the function $f(z)$, each one likewise being counted as often as its order requires. The result given by Eq. 282 states that the contour integral formed for the ratio f'/f, when multiplied by $1/2\pi j$, is equal to the total number of enclosed zeros of the function $f(z)$ diminished by the total number of enclosed poles, each of these being counted as often as its order requires.

An application of this result to a special case is of particular interest. Here the function $f(z)$ is assumed to be the finite polynomial

$$f(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 \qquad [283]$$

Then

$$\frac{f'(z)}{f(z)} = \frac{nz^{n-1} + (n-1)a_{n-1}z^{n-2} + \cdots + a_1}{z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0} \qquad [284]$$

Dividing the denominator into the numerator by long division, one obtains a series of the form

$$\frac{f'(z)}{f(z)} = \frac{n}{z} + \frac{b_2}{z^2} + \frac{b_3}{z^3} + \cdots \qquad [285]$$

which converges only for z-values lying outside a circle enclosing all the zeros of the function $f(z)$. In the region *outside* this circle it represents the Laurent expansion for the function $f'(z)/f(z)$, whence the integer n is recognized to be equal to the sum of the residues of this function in its poles, all of which are located within the circle. Consequently, if C denotes this circular contour, then

$$\frac{1}{2\pi j}\oint_C \frac{f'(z)}{f(z)} = n \qquad [286]$$

Inasmuch as the function $f(z)$, given by Eq. 283, has no poles inside the contour C, the conclusion follows that the highest power n of the polynomial 283 equals the total number of roots of the algebraic equation $f(z) = 0$. This is the familiar *fundamental law of algebra.*

20. A METHOD FOR THE DETECTION OF ZEROS WITHIN A GIVEN REGION

Another useful application of the result stated by Eq. 282 in Art. 19 is the following. If a given function $w = f(z)$ is known to have only zeros within a region enclosed by the contour C, then

$$\frac{1}{2\pi j}\oint_C \frac{f'(z)}{f(z)}\,dz = N \qquad [287]$$

equals the number of enclosed zeros. According to the principles of conformal mapping, the contour C in the z-plane determines an equivalent contour in the w-plane. The integral 287 may correspondingly be replaced by

$$\frac{1}{2\pi j}\oint_D \frac{dw}{w} = N \qquad [288]$$

in which D is a closed contour in the w-plane corresponding to the path C in the z-plane. From the discussion of the logarithm function in Art. 17, it is recognized that the integer N in Eq. 288 must equal the number of times that the contour D in the w-plane encircles the origin. If this

contour encircles the origin more than once — that is, if $N > 1$ — it is clear that the inverse of the function $w = f(z)$, namely, $z = \phi(w)$, is multivalued, and that its Riemann surface in the w-plane has such a structure that a path corresponding to C in the z-plane closes upon itself after N complete circuits around the origin in the w-plane.

The contour C may be located anywhere in the z-plane. The conclusion to be drawn is that the number of zeros of $f(z)$ enclosed by C (each being counted as often as its order requires) is given by the number of times that the corresponding closed contour D encircles the origin in the w-plane. If the closed contour D (which may, when plotted in the simple w-plane, be any closed curve with or without crossover points) does not encircle the origin in the w-plane, the given function $f(z)$ cannot possess any zeros within the region enclosed by C. Hence if it is to be established whether a given function $f(z)$ possesses zeros within a stated region, the question may be answered through mapping in the w-plane the locus corresponding to a boundary enclosing the stated region in the z-plane, and noting whether this closed locus in the w-plane does or does not encircle the origin.

It is, of course, necessary to know that $f(z)$ does not have poles within the region in question. The method is evidently also applicable to the determination of the presence of poles within a given region if the function is known to have no zeros there. In either case, obviously, neither zeros nor poles should lie upon the contour C.

In applying the method to practical problems it is helpful to note the significance of several detailed characteristics which the contour D may exhibit. The contour C is assumed to enclose a simply connected region in the z-plane. If this region contains neither zeros nor poles, the corresponding contour D in the w-plane, as stated above, does not enclose the origin; this characteristic of the contour D, however, may not always be immediately evident. The inverse function $z = \phi(w)$, except in trivial cases, is multivalued, and hence the region enclosed by the contour D may have overlapping portions which actually, of course, lie on separate leaves of the associated Riemann surface. Thus, if the region enclosed by the contour C contains saddle points (see Art. 14), the contour D encloses the corresponding branch points, and hence makes several complete circuitations about portions of the region enclosed by it. One readily appreciates, therefore, that the contour D may be a rather tortuous path even though the contour C is a simple one. It is possible for the region enclosed by the contour D to *surround* the origin completely and yet not *contain* the origin.

The closed contour C is ordinarily assumed to be traversed in the counterclockwise direction, and the region enclosed by it is taken to be that on the left of this contour. The corresponding direction of traversal

for the contour D may always be established through considering any two neighboring points on C and determining the corresponding neighboring points on D. According to the principles of conformal mapping, the region enclosed by D lies to the left of this contour, but this observation alone is frequently insufficient to enable one to tell by inspection whether the enclosed region contains the origin, and if so, how many times the contour D encircles the origin.

A method for the correct evaluation of this situation is the following. One imagines a radius vector extending from the origin of the w-plane to a variable point w on D corresponding to the variable point z on C. As z traverses C and w traverses D, this radius vector changes in length and in its angular position. For a complete traversal of z around C, the net change in the angle of the radius vector must obviously equal an integer number of 2π radians (a change is positive if it corresponds to rotation in the counterclockwise direction). The value of this integer equals the number of times that the contour D encircles the origin.

If the function $w = f(z)$ is analytic within the region enclosed by C, this integer cannot be negative, but it can become negative if the region also contains poles. For example, suppose $f(z) = 1/z$ and let C be any contour enclosing the origin in the z-plane. Then it is clear that the integer in question equals -1. One may say that the contour D in this case encloses the point at infinity in the w-plane. If C is the unit circle, in this simple example D is likewise a unit circle. The enclosed region in the z-plane is that within the unit circle; the enclosed region in the w-plane is that lying *outside* the unit circle. The requirement that the function $f(z)$ be known to have no poles (or other singularities) within the region enclosed by C is again recognized as being necessary, since the presence of a pole cancels the effect of a zero as far as the net value of the integer is concerned.

21. The principle of the maximum modulus; Rouché's theorem and Schwarz's lemma

To continue the considerations of the previous article, if the function $w = f(z)$ is analytic on C and within the region enclosed by it, and if on this contour $|f(z)| \leqq M$ (some real positive quantity), it follows that the contour D and its enclosed region must lie within (or be tangent to) a circle of radius M concentric with the origin of the w-plane. This fact is clear from the consideration that the region enclosed by D cannot extend beyond a boundary formed by those confluent segments of the contour D which are farthest (but still at a finite distance) from the origin; otherwise the point at infinity would be contained within the region, thus contradicting the assumption that $f(z)$ is analytic within C. Since

every point within the region enclosed by C yields a point in the w-plane which lies within the region enclosed by D, it follows that $|f(z)| \leqq M$ for all points within C. Moreover, the equality sign in this relationship holds only if the region enclosed by D consists of a single point, in which case it holds identically, for then $f(z)$ must reduce to the constant M.

This result is known as the *principle of the maximum modulus*. It may readily be demonstrated analytically with the help of Cauchy's integral formula, Eq. 80. The closed contour S (upon and within which $f(z)$ is analytic) is identified with a circle of radius ρ about some finite point z. Then

$$\zeta - z = \rho e^{j\theta} \tag{289}$$

$$f(\zeta) = f(z + \rho e^{j\theta}) \tag{290}$$

and

$$\frac{d\zeta}{\zeta - z} = j\, d\theta \tag{291}$$

Hence one has

$$f(z) = \frac{1}{2\pi}\int_0^{2\pi} f(z + \rho e^{j\theta})\, d\theta \tag{292}$$

from which the value of the function at the center of the circular region is seen to equal the arithmetic mean of its values on the boundary.

Since the mean of a set of complex values must be less than or at most equal to the mean formed from the magnitudes of these values, it follows that the magnitude of $f(z)$ must surely be less than or at most equal to the largest value which the magnitude of the function assumes on the circular boundary. That is, if M is the maximum value of $|f(\zeta)|$ on the boundary, $|f(z)| \leqq M$.

This result may be generalized to the extent that the boundary need not be circular and the point z may be any internal point. For if the result were not also true in this more general case — that is, if the maximum of the absolute value of the function did not occur on the boundary but at an internal point — by applying the specialized result to an appropriately chosen small circle about this internal point, one would clearly encounter a contradiction. This line of thought also shows at once that the equality of $|f(z)|$ and M can hold only if it holds identically, that is, if $f(z)$ equals the constant M.

If the function $f(z)$ has no zeros (as well as no poles) upon or within the closed contour, by applying the same reasoning to the reciprocal function $1/f(z)$, one recognizes that *both* the minimum and the maximum values of $|f(z)|$ on the boundary are minima and maxima for the enclosed region.

An alternative proof, which is collaterally interesting, begins with the result expressed by Eq. 287 of the previous article. The function $f(z)$ is analytic upon C and within the region enclosed by this contour. A second function $g(z)$, in addition to satisfying the same analyticity conditions, fulfills the relation

$$|g(z)| < |f(z)| \tag{293}$$

on the boundary C. On this boundary one, therefore, has

$$\left|\frac{g(z)}{f(z)}\right| < 1 \tag{294}$$

and hence the closed contour for the function

$$w(z) = 1 + \frac{g(z)}{f(z)} \tag{295}$$

in the w-plane, corresponding to the contour C in the z-plane, clearly cannot enclose the point $w = 0$. According to the discussion in Art. 20, therefore, one has

$$\oint_C \frac{dw}{w} = 0 \tag{296}$$

Using Eq. 295, one now finds

$$\frac{dw}{w} = \frac{f \cdot d(f + g) - (f + g) \cdot df}{f(f + g)} = \frac{d(f + g)}{f + g} - \frac{df}{f} \tag{297}$$

Hence in view of Eq. 296, the result expressed by Eq. 287 gives

$$\frac{1}{2\pi j}\oint_C \frac{d(f + g)}{f + g} = \frac{1}{2\pi j}\oint_C \frac{df}{f} = N \tag{298}$$

from which it may be concluded that, under th conditions stated above, the function $f(z) + g(z)$ has the same number of zeros within the region enclosed by C as does the function $f(z)$.

In terms of this result (known as *Rouché's theorem*), the principle of the maximum modulus is easily proved. Suppose the maximum of $|f(z)|$ did not occur on the boundary C but at some internal point z_0. Then on the boundary one would have

$$\frac{1}{|f(z_0)|} < \frac{1}{|f(z)|} \tag{299}$$

and hence one could conclude that the function $[1/f(z_0) - 1/f(z)]$ has the same number of zeros within the region enclosed by C as does the function $1/f(z)$. The latter has no zeros within this region because $f(z)$ is analytic there, and the function $[1/f(z_0) - 1/f(z)]$ has at least one zero

within the region, namely the one for $z = z_0$. The supposition is, therefore, untenable.

A more specialized result is obtained if one considers the function $f(z)$ to be analytic within a circle of radius R about the origin and to equal zero at the origin. It then possesses the Maclaurin series

$$f(z) = a_1 z + a_2 z^2 + a_3 z^3 + \cdots \qquad [300]$$

from which it is clear that the function

$$\phi(z) = \frac{f(z)}{z} \qquad [301]$$

likewise is analytic within the circle of radius R. If on the circular boundary it is known that

$$|f(z)| \leqq M \qquad [302]$$

then it follows that there

$$|\phi(z)| \leqq \frac{M}{R} \qquad [303]$$

and, according to the principle of the maximum modulus, that the magnitude of $\phi(z)$ is less than (or at most equal to) M/R for all points within the circle. Hence one has the result that

$$|f(z)| \leqq \frac{|z|}{R} \cdot M \qquad [304]$$

for all points within the circle, and the equality can hold only if it holds identically, in which case $f(z)$ reduces to a complex constant with the magnitude M/R, multiplied by z.

This particular result is known as *Schwarz's lemma*. It has numerous practical applications in problems involving conformal mapping.*

22. Some useful correlations with potential theory; Poisson's integrals and Hilbert transforms

In Art. 5 it is pointed out that the real and imaginary parts u and v of a function of a complex variable may be interpreted physically as being two-dimensional potential functions. Any two functions of the variables x and y which satisfy the Cauchy-Riemann Eqs. 12 and 13 are said to be conjugate potential functions. Since Eqs. 12 and 13 become interchanged when u is replaced by v and v by $-u$, it is clear that the identities of these two functions become interchanged when the algebraic

*An application of this sort is discussed in Art. 27.

sign of one of them is reversed. In the present article the properties of these functions are investigated in more detail.

The point of departure in these discussions is the Cauchy integral formula, Eq. 80. A given function $f(z)$ is assumed to be regular within the region enclosed by the circle shown in Fig. 21 as well as on this boundary. If ζ denotes any point on the circle and z is any internal point,

$$f(z) = \frac{1}{2\pi j}\oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta \qquad [305]$$

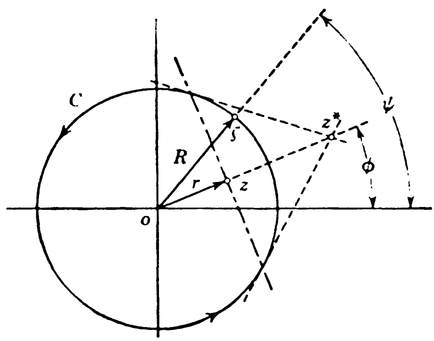


FIG. 21. Change of variable relevant to the derivation of Poisson's integrals.

If, in this integral, z is replaced by some point z^* external to the circle, the integrand is regular for all points enclosed by the contour C, and according to the residue theorem or Cauchy's integral law, the value of the integral is zero. That is,

$$0 = \frac{1}{2\pi j}\oint_C \frac{f(\zeta)}{\zeta - z^*}\,d\zeta \qquad [306]$$

The point z^* is now so chosen that

$$z^* = \frac{|\zeta|^2}{\bar{z}} = \frac{\zeta\bar{\zeta}}{\bar{z}} \qquad [307]$$

in which the bar indicates the conjugate value.* Using this relation, one finds

$$\frac{1}{\zeta - z} \pm \frac{1}{\zeta - z^*} = \frac{1}{\zeta}\left\{\frac{\zeta}{\zeta - z} \mp \frac{\bar{\zeta}}{\bar{\zeta} - \bar{z}} \pm 1\right\} \qquad [308]$$

*The point z^* thus determined is called the *image* of the point z with respect to the circle. The two points z and z^* are geometrically related as shown in Fig. 21. This item is discussed in Art. 24.

Addition or subtraction of Eqs. 305 and 306 is then seen to yield

$$f(z) = \pm f(0) + \frac{1}{2\pi j}\oint_C \left\{\frac{\zeta}{\zeta - z} \mp \frac{\bar{\zeta}}{\bar{\zeta} - \bar{z}}\right\} f(\zeta)\frac{d\zeta}{\zeta} \qquad [309]$$

in which Eq. 305 is also used to obtain the particular relation

$$f(0) = \frac{1}{2\pi j}\oint_C f(\zeta)\frac{d\zeta}{\zeta} \qquad [310]$$

Since, according to the notation given in Fig. 21,

$$z = re^{j\phi} \qquad [311]$$

and

$$\zeta = Re^{j\psi} \qquad [312]$$

one has

$$\frac{\zeta}{\zeta - z} = \frac{Re^{j\psi}}{Re^{j\psi} - re^{j\phi}} = \frac{R^2 - rRe^{-j(\psi-\phi)}}{R^2 + r^2 - 2rR\cos(\psi - \phi)} \qquad [313]$$

Hence

$$\frac{\zeta}{\zeta - z} - \frac{\bar{\zeta}}{\bar{\zeta} - \bar{z}} = \frac{-j2rR\sin(\psi - \phi)}{R^2 + r^2 - 2rR\cos(\psi - \phi)} \qquad [314]$$

and

$$\frac{\zeta}{\zeta-z}+\frac{\bar{\zeta}}{\bar{\zeta}-\bar{z}}=\frac{2R^2-2rR\cos(\psi-\phi)}{R^2+r^2-2rR\cos(\psi-\phi)}=1+\frac{R^2-r^2}{R^2+r^2-2rR\cos(\psi-\phi)} \qquad [315]$$

Equation 312 is used to obtain

$$\frac{d\zeta}{\zeta} = j\,d\psi \qquad [316]$$

Substituting these results into Eq. 309, one obtains the following integral representations:

$$f(z) = f(0) + \frac{1}{j\pi}\int_0^{2\pi}\frac{rR\sin(\psi-\phi)}{R^2+r^2-2rR\cos(\psi-\phi)}f(R,\psi)\,d\psi \qquad [317]$$

and

$$f(z) = \frac{1}{2\pi}\int_0^{2\pi}\frac{R^2 - r^2}{R^2 + r^2 - 2rR\cos(\psi - \phi)}f(R,\psi)\,d\psi \qquad [318]$$

It is useful to observe that the quantity $rR\sin(\psi - \phi)$ represents the area of the parallelogram determined by the vectors ζ and z and that

$$R^2 + r^2 - 2rR\cos(\psi - \phi) = |\zeta - z|^2 \qquad [319]$$

is the square of the distance between the points ζ and z. It may, therefore, be concluded that the formulas given by Eqs. 317 and 318 are independent of the location of the center of the circle C, although in the above derivation the center is for convenience chosen to coincide with the origin of the z-plane.

If the function $f(z)$ is written more explicitly as

$$f(z) = u(r,\phi) + jv(r,\phi) \tag{320}$$

and Eqs. 317 and 318 are separated into their real and imaginary parts, one obtains the following two pairs of relations respectively:

$$u(r,\phi) = u(0) + \frac{1}{\pi}\int_0^{2\pi} \frac{rR\sin(\psi - \phi)}{|\zeta - z|^2} v(R,\psi)\,d\psi \tag{321}$$

$$v(r,\phi) = v(0) - \frac{1}{\pi}\int_0^{2\pi} \frac{rR\sin(\psi - \phi)}{|\zeta - z|^2} u(R,\psi)\,d\psi \tag{322}$$

and

$$u(r,\phi) = \frac{1}{2\pi}\int_0^{2\pi} \frac{R^2 - r^2}{|\zeta - z|^2} u(R,\psi)\,d\psi \tag{323}$$

$$v(r,\phi) = \frac{1}{2\pi}\int_0^{2\pi} \frac{R^2 - r^2}{|\zeta - z|^2} v(R,\psi)\,d\psi \tag{324}$$

in which the relation 319 is used for the sake of abbreviation. It is observed that the formulas in each pair become interchanged when u is replaced by v and v by $-u$.

The results expressed by Eqs. 321 and 322 show that the real and imaginary parts, u and v, of a function of a complex variable are, except for an additive constant, explicitly related to each other. The real part determines the imaginary part, or vice versa. Hence, given either part, the corresponding complex function may be found. The formulas 321 and 322 are said to yield the conjugate potential function to any given potential function or to transform one such function into its conjugate mate. In this sense they are sometimes referred to as a pair of transforms.

The integrals 323 and 324 likewise form a pair, but they do not express one potential function in terms of the other. Instead, they yield the real and imaginary parts of a complex function at any point within the given circle in terms of their respective values on the boundary. If a potential function is known to be regular for all points on a circular boundary and within the enclosed region, it is there uniquely determined in terms of its boundary values by means of the formula 323 or 324.

For $r = 0$, these integrals yield

$$u(0) = \frac{1}{2\pi}\int_0^{2\pi} u(R,\psi)\,d\psi \tag{325}$$

and

$$v(0) = \frac{1}{2\pi}\int_0^{2\pi} v(R,\psi)\, d\psi \qquad [326]$$

These results state that the value of either potential function at the center of the circle is equal to the *average* of its values on the circular boundary. As a consequence it follows that the largest and smallest of the values which u and v assume throughout a circular regularity region must occur on the boundary of that region, for if such an extremum occurred at some internal point, the conditions stated by Eqs. 325 and 326 could not be fulfilled for a small circle with its center at this point and its boundary within the original one.

These conclusions yield a useful theorem to the effect that if a function of a complex variable is regular over a given region, the maximum and minimum values of its real and imaginary parts must for that region occur on the boundary.* The region evidently is not restricted to be circular in form since the above reasoning is equally applicable when the boundary of the regularity region is arbitrary.

The formulas 321, 322, 323, and 324 are known as *Poisson's integrals.* By a combination of Eqs. 322 and 323 a further result of practical utility is obtained. The first step is to form

$$u(r,\phi) + jv(r,\phi)$$
$$= jv(0) + \frac{1}{2\pi}\int_0^{2\pi} \frac{R^2 - r^2 - j2rR\sin(\psi - \phi)}{R^2 + r^2 - 2rR\cos(\psi - \phi)}\, u(R,\psi)\, d\psi \qquad [327]$$

From Eqs. 311 and 312 it is seen that

$$\frac{\zeta + z}{\zeta - z} = \frac{Re^{j\psi} + re^{j\phi}}{Re^{j\psi} - re^{j\phi}} = \frac{R^2 - r^2 - j2rR\sin(\psi - \phi)}{R^2 + r^2 - 2rR\cos(\psi - \phi)} \qquad [328]$$

so that Eq. 327 may be written

$$f(r,\phi) = jv(0) + \frac{1}{2\pi}\int_0^{2\pi} \frac{\zeta + z}{\zeta - z}\, u(R,\psi)\, d\psi \qquad [329]$$

By means of the series representation

$$\frac{\zeta + z}{\zeta - z} = 1 + 2\left(\frac{z}{\zeta} + \frac{z^2}{\zeta^2} + \frac{z^3}{\zeta^3} + \cdots\right) \qquad [330]$$

*This result, although similar to that stated by the principle of the maximum modulus, should not be confused with the latter. As pointed out in the previous article, the modulus $\sqrt{u^2 + v^2}$ attains its minimum as well as its maximum value on the boundary of the region of analyticity only if the function has no zeros within this region. That is, if the function has zeros there, the modulus attains its maximum value, but not its minimum value, on the boundary.

and Eq. 325, this result may be rewritten in the form

$$f(r,\phi) = f(0) + \frac{1}{\pi}\int_0^{2\pi}\left(\frac{z}{\zeta} + \frac{z^2}{\zeta^2} + \cdots\right)u(R,\psi)\,d\psi \qquad [331]$$

The series in Eq. 330 is uniformly convergent for $\left|\frac{z}{\zeta}\right| < 1$, so that the integration in Eq. 331 may be carried out term by term. When this is done, with

$$\alpha_n = \frac{1}{\pi}\int_0^{2\pi} u(R,\psi)e^{-jn\psi}\,d\psi \qquad [332]$$

it is seen that the result given by Eq. 329 may be expressed in the alternative form

$$f(r,\phi) = f(0) + \sum_{n=1}^{\infty}\left(\frac{r}{R}\right)^n \alpha_n e^{jn\phi} \qquad [333]$$

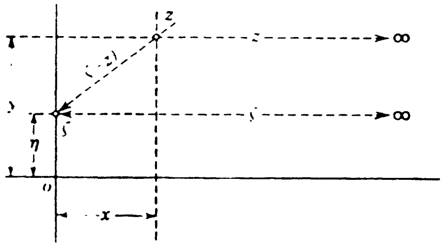


FIG. 22. Limiting process which converts Poisson's integrals into Hilbert transforms.

By means of either Eq. 329 or Eq. 333, the complex function is determined within the circular region in terms of the values of its real part on the boundary.

Other forms for some of these results, more appropriate to the conditions encountered in electric circuit theory, are obtained through assuming that the circle in the preceding derivations is an extremely large one lying in the right half of the z-plane, tangent to the y-axis at the origin. If this circle is imagined to become infinite in diameter, the entire right half plane will constitute the enclosed region and the imaginary axis (or y-axis) will become the boundary, which then closes upon itself by passing through the point at infinity. The manner in which the integrals 317 and 318 are to be interpreted in this limiting case is clarified somewhat when the situation is pictured in Fig. 22. The center of the circle of Fig. 21 is imagined to lie on the real axis at a point infinitely far to the right. The point ζ, which in Fig. 21 is a point on the circular boundary, becomes the point $j\eta$ on the imaginary axis, and the point z is character-

ized in terms of the variables x and y of the plane of Fig. 22 as the point $x + jy$.

In order to evaluate the forms which the integrals 317 and 318 assume when this limiting process is carried out, one should observe that the appropriate transition may be indicated symbolically as follows:

$$r \sin (\psi - \phi) \rightarrow (y - \eta) \qquad [334]$$

$$R^2 - r^2 \rightarrow R^2 - (R - x)^2 \rightarrow 2Rx \qquad [335]$$

and

$$|\zeta - z|^2 \rightarrow x^2 + (y - \eta)^2 \qquad [336]$$

whereas

$$R\, d\psi \rightarrow -d\eta \qquad [337]$$

in which the minus sign appears because the positive direction of traversal of the circular boundary in Fig. 21 is counterclockwise and in Fig. 22 this corresponds to the negative direction for the y-axis. Finally it should be observed that the limits of integration corresponding to a counterclockwise traversal of the circle are from $+\infty$ to $-\infty$, or from $-\infty$ to $+\infty$ if the algebraic sign of the integral is reversed.

When these substitutions are made and $f(z)$ is assumed to vanish at infinity,* the integrals 317 and 318 become in the limit

$$f(z) = u(x,y) + jv(x,y) = \frac{1}{j\pi} \int_{-\infty}^{\infty} \frac{(y - \eta) f(j\eta)}{x^2 + (y - \eta)^2}\, d\eta \qquad [338]$$

and

$$f(z) = u(x,y) + jv(x,y) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x f(j\eta)}{x^2 + (y - \eta)^2}\, d\eta \qquad [339]$$

The transformation of the Poisson integrals, which in their usual form apply to a circularly bounded region, to the forms given by Eqs. 338 and 339, which apply to the right half plane, may be carried out in a somewhat more satisfactory manner (than the heuristic one just given) by use of a linear fractional transformation (discussed in Art. 24) which effects the mapping of the interior of a circle about the origin upon the right half plane.

From Art. 24, one finds that the transformation which reads

$$z = \frac{z' - a}{z' + a} \qquad [340]$$

transforms the interior of the unit circle in the z-plane into the right half

*The validity of the integral in Eq. 338 is restricted to this condition. The term $f(0)$ in Eq. 317 becomes $f(\infty)$ and drops out.

of the z'-plane, with the origin of the z-plane corresponding to the point $z' = a$ on the real axis of the z'-plane. As the unit circle in the z-plane is traversed in the counterclockwise direction, the imaginary axis in the z'-plane is traversed throughout its entire extent from $+j\infty$ to $-j\infty$ (that is, in the negative direction).

If z is some point within the unit circle, z' is a point within the right half plane. To a point ζ on the unit circle in the z-plane there corresponds a point ζ' on the imaginary axis of the z'-plane which likewise is determined from the transformation 340, that is,

$$\zeta = \frac{\zeta' - a}{\zeta' + a} \qquad [341]$$

The desired transformation of the Poisson integrals is carried out through returning to Eq. 309 and introducing there the change of variable expressed by Eqs. 340 and 341. One finds from a simple calculation that

$$\frac{\zeta}{\zeta - z} = \frac{(\zeta' - a)(z' + a)}{2a(\zeta' - z')} \qquad [342]$$

and

$$\frac{d\zeta}{\zeta} = \frac{2a\, d\zeta'}{(\zeta' - a)(\zeta' + a)} \qquad [343]$$

The integrand appearing in the integral of Eq. 309, therefore, becomes

$$\left\{\frac{(\zeta' - a)(z' + a)}{\zeta' - z'} \mp \frac{(\bar{\zeta}' - a)(\bar{z}' + a)}{\bar{\zeta}' - \bar{z}'}\right\} \frac{f(\zeta')\, d\zeta'}{(\zeta' - a)(\zeta' + a)} \qquad [344]$$

It should now be observed that one wishes to have the origin of the z-plane correspond to a point in the z'-plane which is infinitely remote from the origin, because the value of the function at the origin in the z-plane (the quantity $f(0)$ in Eq. 309) is then carried over into the value of the function at infinity in the z'-plane. Under the assumption that the function vanishes at infinity (which must, of course, be met by any physical problem to which the resulting formulas are applied), the term involving $f(0)$ in Eq. 309 then drops out.

Although it is not possible to determine a linear fractional transformation which maps the interior of a circle about the origin upon the right half plane in such a way as to make the origin or the center of the circle correspond to the point at infinity (this would require an infinitely large value for the quantity a in Eq. 340), the desired end may be achieved in the present problem by considering the limiting process indicated by $a \to \infty$ to be applied to the expression 344. The result reads

$$\left\{\frac{1}{\zeta' - z'} \mp \frac{1}{\bar{\zeta}' - \bar{z}'}\right\} f(\zeta')\, d\zeta' \qquad [345]$$

Substituting this result for the integrand appearing in Eq. 309, discarding the term $f(0)$ for reasons already stated, and noting that the algebraic sign of the integral is reversed if the integration is extended over the imaginary axis of the z'-plane in the positive direction, one has after dropping the primes on the quantities z and ζ

$$f(z) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left\{\frac{1}{z-\zeta} \mp \frac{1}{\bar{z}-\bar{\zeta}}\right\} f(\zeta)\,d\zeta \qquad [346]$$

which is the desired result. It is readily brought into the more explicit form given by Eqs. 338 and 339 through writing $z = x + jy$ for any fixed point in the right half plane and $\zeta = j\eta$ for the variable point of integration along the boundary or imaginary axis. The two algebraic signs appearing in the integrand then yield respectively the results

$$f(z) = \frac{1}{j\pi}\int_{-\infty}^{\infty} \frac{(y-\eta)f(j\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [347]$$

and

$$f(z) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{xf(j\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [348]$$

which agree with Eqs. 338 and 339.

Separating real and imaginary parts yields the pairs

$$u(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{(y-\eta)v(0,\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [349]$$

$$v(x,y) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{(y-\eta)u(0,\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [350]$$

and

$$u(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{xu(0,\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [351]$$

$$v(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{xv(0,\eta)}{x^2+(y-\eta)^2}\,d\eta \qquad [352]$$

The last two integrals determine the potential functions in the right half plane in terms of their respective values on the imaginary axis, which is regarded as the boundary of the right half plane. The individual relations in the pair 349 and 350, on the other hand, may be used to determine either of the potential functions in the right half plane in terms of the boundary values of the conjugate function. A potential function along the *boundary* is given in terms of the conjugate function by either of the

integrals 349 or 350 for $x = 0$. One thus obtains the pair of relations

$$u(y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{v(\eta)\,d\eta}{y - \eta} \qquad [353]$$

and

$$v(y) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(\eta)\,d\eta}{y - \eta} \qquad [354]$$

which are known as *Hilbert transforms.*

Because of the singularity of the integrand at the point $\eta = y$, the value of either of these integrals, in the ordinary sense, does not exist. It is, however, possible to overcome this difficulty* by the definition of a particular process of evaluation yielding the so-called *Cauchy principal value*. This value is obtained through approaching the point $\eta = y$ symmetrically from both sides, as indicated for the integral 353 in the following expression:

$$\int_{-\infty}^{\infty} \frac{v(\eta)\,d\eta}{y - \eta} = \operatorname*{limit}_{\epsilon \to 0}\left[\int_{-\infty}^{y-\epsilon} \frac{v(\eta)\,d\eta}{y - \eta} + \int_{y+\epsilon}^{\infty} \frac{v(\eta)\,d\eta}{y - \eta}\right] \qquad [355]$$

In order to see that this limit has a definite value, one may represent the function $v(\eta)$ as

$$v(\eta) = v(y) + (\eta - y)\cdot p(y,\eta) \qquad [356]$$

in which $p(y,\eta)$ is regular in the vicinity of the point $\eta = y$. Then

$$\int_{-\infty}^{\infty} \frac{v(\eta)\,d\eta}{y - \eta} = v(y)\int_{-\infty}^{\infty} \frac{d\eta}{y - \eta} - \int_{-\infty}^{\infty} p(y,\eta)\,d\eta \qquad [357]$$

The principal value of the first of the two integrals on the right-hand side is obtained through the following steps:

$$\int_{-\infty}^{y-\epsilon} \frac{d\eta}{y - \eta} + \int_{y+\epsilon}^{\infty} \frac{d\eta}{y - \eta} = -\Big[\ln(y - \eta)\Big]_{-\infty}^{y-\epsilon} - \Big[\ln(y - \eta)\Big]_{y+\epsilon}^{\infty} \qquad [358]$$

When this is written in the form

$$\Big[\ln(y - \eta)\Big]_{y-\epsilon}^{y+\epsilon} - \Big[\ln(y - \eta)\Big]_{-\infty}^{\infty} \qquad [359]$$

the limiting process is found to yield

$$\operatorname*{limit}_{\epsilon \to 0}\left[\ln\left(\frac{-\epsilon}{\epsilon}\right)\right] - \operatorname*{limit}_{\eta \to \infty}\left[\ln\left(\frac{y - \eta}{y + \eta}\right)\right] = 0 \qquad [360]$$

*See E. C. Titchmarsh, "Conjugate Trigonometric Integrals," *Proceedings of the London Mathematical Society*, 2nd series, *24* (1926) 109–130; "On Conjugate Functions," *ibidem*, *29* (1929) 49–81.

Hence only the second integral on the right-hand side of Eq. 357 remains.

According to Eq. 356, the integrand $p(y,\eta)$ of this second integral is observed to decrease no faster than $1/\eta$ for large values of η regardless of whether $v(\eta)$ vanishes or remains finite for $\eta \to \infty$. Although it is true that the integral from zero to infinity of a function which decreases no faster than $1/\eta$ for large η does not remain finite, it is essential in the present instance to observe that if the limits of the integral are from $-\infty$ to ∞, and the integrand is an odd function about the point $\eta = \infty$, one can appreciate in a general way how a finite value can result by noting that the contributions to the integral on the positive and negative sides of the point ∞ have opposite algebraic signs. This is the same sort of reasoning as that involved in deriving the finite principal value 355.

By reference to Eq. 356 it is seen that if $v(\eta)$ vanishes for $\eta \to \infty$, then $p(y,\eta)$ is an odd function about the point $\eta = \infty$; but if $v(\eta)$ remains finite in this vicinity, $p(y,\eta)$ is odd about $\eta = \infty$ only if $v(\eta)$ is an even function; that is, if $v(-\eta) = v(\eta)$. Hence one is led to the conclusion that the given potential function $v(\eta)$ either must vanish for $\eta \to \infty$ or must be an even function of η. Inasmuch as an odd function about ∞ must necessarily be either zero or infinite at infinity, it is sufficient to state merely that the potential function $v(\eta)$ shall remain regular in the vicinity of $\eta = \infty$.* Since all the above discussion applies equally to the evaluation of the integral 354, the same final comment applies also to the conjugate function $u(\eta)$.

It is thus seen that the transforms 353 and 354 may be written in the alternative form

$$u(y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{v(\eta) - v(y)}{y - \eta}\,d\eta \qquad [361]$$

$$v(y) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(\eta) - u(y)}{y - \eta}\,d\eta \qquad [362]$$

Here the functions $v(\eta) - v(y)$ and $u(\eta) - u(y)$ vanish in the vicinity of the point $\eta = y$ at least as strongly as the factor $(y - \eta)$, so that there can be no question as to the finiteness of the integrals so far as the presence of the factor $(y - \eta)$ in the denominator of the integrand is concerned.

In the evaluation of the integrals 351 and 352 for the boundary $x = 0$, a difficulty arises due to the fact that the integrand vanishes everywhere with the possible exception of the point $\eta = y$. Yet it must be true that, in the limit $x \to 0$, the integrals yield the values $u(0,y)$ and $v(0,y)$. That they do so, may be shown in the following way.

*For a more thorough discussion of the question of convergence of the Hilbert transforms the reader is referred to the article by E. C. Titchmarsh already cited.

Since the only possible contribution to the value of the integral in the limit $x \to 0$ must come from the immediate vicinity of the point $\eta = y$, it is clear that for example, the function $u(0,\eta)$, in Eq. 351, may be replaced by its value $u(0,y)$ at the point $\eta = y$, and since it then has nothing to do with the process of integration it may be placed in front of the integral sign. It then remains to show that

$$u(0,y) = \frac{u(0,y)}{\pi}\left\{\operatorname*{limit}_{x\to 0}\int_{-\infty}^{\infty}\frac{x\,d\eta}{x^2+(y-\eta)^2}\right\} \qquad [363]$$

or that

$$\operatorname*{limit}_{x\to 0}\left\{\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x\,d\eta}{x^2+(y-\eta)^2}\right\} = \operatorname*{limit}_{x\to 0}\left[\frac{1}{\pi}\tan^{-1}\left(\frac{\eta-y}{x}\right)\right]_{\eta=-\infty}^{\eta=\infty} = 1 \qquad [364]$$

which is evidently true.

To return to the integrals 353 and 354 (or the equivalent ones given by Eqs. 361 and 362, it is useful to observe that if $v(\eta)$ is an odd function of η, then $u(\eta)$ is an even function, and vice versa. To see this, one may write the integral 353, for example, as

$$u(y) = \frac{1}{\pi}\int_{-\infty}^{0}\frac{v(\eta)\,d\eta}{y-\eta} + \frac{1}{\pi}\int_{0}^{\infty}\frac{v(\eta)\,d\eta}{y-\eta} \qquad [365]$$

or

$$u(y) = -\frac{1}{\pi}\int_{0}^{-\infty}\frac{v(\eta)\,d\eta}{y-\eta} + \frac{1}{\pi}\int_{0}^{\infty}\frac{v(\eta)\,d\eta}{y-\eta} \qquad [366]$$

If $v(\eta)$ is assumed to be *odd*, replacing the variable of integration η in the first of the above two integrals by $-\eta$ shows that

$$-\frac{1}{\pi}\int_{0}^{-\infty}\frac{v(\eta)\,d\eta}{y-\eta} = -\frac{1}{\pi}\int_{0}^{\infty}\frac{v(\eta)\,d\eta}{y+\eta} \qquad [367]$$

Hence Eq. 366 becomes

$$u(y) = \frac{1}{\pi}\int_{0}^{\infty}\left(\frac{1}{y-\eta} - \frac{1}{y+\eta}\right)v(\eta)\,d\eta \qquad [368]$$

or

$$u(y) = -\frac{2}{\pi}\int_{0}^{\infty}\frac{\eta v(\eta)}{\eta^2-y^2}\,d\eta \qquad [369]$$

From this form it is immediately evident that $u(y)$ is an *even* function of y.

By means of an entirely parallel process, one may transform the

integral 354 with $u(\eta)$ assumed to be *even* and obtain

$$v(y) = \frac{2y}{\pi}\int_0^\infty \frac{u(\eta)}{\eta^2 - y^2}\,d\eta \qquad [370]$$

from which it is apparent that $v(y)$ is an *odd* function of y.

It is to be observed that the transforms 369 and 370 are no longer general but apply only to cases in which u is even and v odd, as is assumed in their derivation. This restricted character of the forms 369 and 370 is recognized at once from the fact that these integrals no longer become interchanged when u is replaced by v and v by $-u$.

One might just as well have derived special forms for the opposite assumptions, namely that u be odd and v even. However, for the rational functions and many others met in practical problems, the real part is even and the imaginary part is odd when regarded as a function of y for $x = 0$, that is, for points on the imaginary axis. The following more detailed discussion pertains to this important special case.

If the same method of transformation is applied to the equivalent forms 361 and 362, the results obtained are

$$u(y) = -\frac{2}{\pi}\int_0^\infty \frac{\eta v(\eta) - y v(y)}{\eta^2 - y^2}\,d\eta \qquad [371]$$

and

$$v(y) = \frac{2y}{\pi}\int_0^\infty \frac{u(\eta) - u(y)}{\eta^2 - y^2}\,d\eta \qquad [372]$$

The last of these is the same as 370 except that $u(\eta) - u(y)$ replaces $u(\eta)$.*

It should be mentioned that the even function $u(y)$ is determined by the integrals 369 or 371 except for an arbitrary additive constant. However, the difference between any two values of $u(y)$ for specified y-values must be uniquely determined. For example, using the integral 369, one finds

$$u(\infty) - u(0) = \frac{2}{\pi}\int_0^\infty v(\eta)\,\frac{d\eta}{\eta} \qquad [373]$$

which is a simple expression for the difference between the values of u at the two extremes in the range of y-values.

This last expression may be put into a practically more useful form through introducing the change of variable indicated by

$$\theta = \ln\left(\frac{\eta}{y_0}\right) \qquad d\theta = \frac{d\eta}{\eta} \qquad [374]$$

*If a given potential function is constant, the conjugate function is zero. Therefore, the conjugate of $u(\eta)$ + constant is the same as the conjugate of $u(\eta)$.

which amounts to introducing a logarithmic scale in place of the linear η- or y-scale and choosing the arbitrary point $\eta = y_0$ as the new origin. If $v(\theta)$ represents the function $v(\eta)$ with respect to the logarithmic θ-scale (that is, $v(\theta)$ is $v(\eta)$ plotted on a logarithmic scale), the relation 373 takes the form

$$u(\infty) - u(0) = \frac{2}{\pi}\int_{-\infty}^{\infty} v(\theta)\, d\theta \tag{375}$$

or

$$\int_{-\infty}^{\infty} v(\theta)\, d\theta = \frac{\pi}{2}\left[u(\infty) - u(0)\right] \tag{376}$$

in which $u(\infty)$ and $u(0)$ still denote the values of u for $y = \infty$ and $y = 0$ respectively.*

This result states that the net area under the curve for the function v, plotted on a logarithmic scale, depends only upon the difference between the values of the corresponding u-function at the extremes of the y-scale. Although this result is only a partial statement of the implicit relation between the conjugate potential functions u and v, its simplicity makes it particularly useful in connection with the practical problem discussed in the references just cited.

A comparable result obtainable from the integral 370 reads

$$\lim_{y \to \infty} \left[y v(y)\right] = -\frac{2}{\pi}\int_0^{\infty} u(\eta)\, d\eta \tag{377}$$

Another useful particular relation between u and v for points on the imaginary axis ($\zeta = j\eta$, $z = jy$) may be obtained from the integral 372. The latter may be rewritten in the form

$$v(y) = \frac{1}{\pi}\int_0^{\infty} \frac{u(\eta) - u(y)}{\frac{1}{2}\left(\frac{\eta}{y} - \frac{y}{\eta}\right)} \cdot \frac{d\eta}{\eta} \tag{378}$$

Introducing the change of variable

$$\theta = \ln\left(\frac{\eta}{y}\right) \qquad d\theta = \frac{d\eta}{\eta} \tag{379}$$

and writing $u(\theta)$ for the function $u(\eta)$ in terms of the new logarithmic variable θ give

$$v(y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(\theta) - u(0)}{\sinh\theta}\, d\theta \tag{380}$$

*This particular form as well as the ones given by Eqs. 377, 380 and 393, were obtained by H. W. Bode. See his U.S. Patent 2,123,178; also "Feedback Amplifier Design," *Bell System Technical Journal*, *19* (July 1940), 421–454.

The integrand in this integral remains regular throughout the entire range $-\infty < \theta < \infty$. At the limits $\theta = \pm\infty$, $u(\theta)$ has the finite values of $u(\eta)$ for $\eta = \infty$ and $\eta = 0$ respectively, whereas $2 \sinh\theta \to \pm e^{|\theta|}$, so that the integrand is seen to behave like $e^{-|\theta|}$ for large positive or negative values of θ. For the vicinity of the point $\theta = 0$,

$$u(\theta) - u(0) \simeq \theta \cdot \left(\frac{du}{d\theta}\right)_{\theta=0} \quad [381]$$

and

$$\sinh\theta \simeq \theta \quad [382]$$

so that the integrand there maintains a finite value.

The integral 380 may be transformed into a more useful form through the method of integration by parts, which reads

$$\int p\,dq = pq - \int q\,dp \quad [383]$$

Letting

$$p = u(\theta) - u(0) \quad [384]$$

and

$$dq = \frac{d\theta}{\sinh\theta} \quad [385]$$

one has

$$dp = \left(\frac{du}{d\theta}\right) d\theta \quad [386]$$

and

$$q = \ln\tanh\frac{\theta}{2} = -\ln\coth\frac{\theta}{2} \quad [387]$$

Thus Eq. 380 becomes

$$v(y) = \frac{1}{\pi}\left[\{u(\theta) - u(0)\} \ln\tanh\frac{|\theta|}{2}\right]_{-\infty}^{\infty} + \frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{du}{d\theta}\right)\ln\coth\frac{|\theta|}{2}\,d\theta \quad [388]$$

in which the absolute value of the variable θ must be used because the integral 380 has a real value and hence the argument of the logarithm must remain positive. Since the quantity $u(\theta) - u(0)$ appearing in the first term of Eq. 388 remains finite at the limits $\theta = \pm\infty$, the value

of this term is zero at both limits. Its values throughout the range $-\infty < \theta < \infty$, moreover, are finite, since at the critical point $\theta = 0$ it has the value

$$\left(\frac{du}{d\theta}\right)_{\theta=0} \cdot \theta \cdot \ln\frac{\theta}{2} \to 0 \qquad [389]$$

This term, therefore, vanishes and there remains

$$v(y) = \frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{du}{d\theta}\right)\ln\coth\frac{|\theta|}{2}\,d\theta \qquad [390]$$

The factor $\ln\coth\frac{|\theta|}{2}$ appearing in the integrand has a logarithmic infinity at the point $\theta = 0$ and is a symmetrical function about this point, dropping off rather rapidly on both sides. Since $\theta = 0$ corresponds to $\eta = y$, it is seen that the value of the function $v(y)$ at any point is largely determined by the slope of the conjugate function $u(\eta)$ at that point when plotted on a logarithmic scale.

This interpretation of the result expressed by Eq. 390 may be clarified through writing

$$\left(\frac{du}{d\theta}\right)_{\theta=0} = \left(\frac{du}{d\eta}\frac{d\eta}{d\theta}\right)_{\eta=y} = y\left(\frac{du}{d\eta}\right)_{\eta=y} = \left(\frac{du}{d\theta}\right)_0 \qquad [391]$$

and noting that*

$$\int_{-\infty}^{\infty}\ln\coth\frac{|\theta|}{2}\,d\theta = 2\int_0^{\infty}\ln\coth\frac{\theta}{2}\,d\theta = \frac{\pi^2}{2} \qquad [392]$$

Then, by adding and subtracting $(du/d\theta)_0$ to $(du/d\theta)$ in the integrand of Eq. 390, one obtains the form

$$v(y) = \frac{\pi}{2}\left(\frac{du}{d\theta}\right)_0 + \frac{1}{\pi}\int_{-\infty}^{\infty}\left\{\left(\frac{du}{d\theta}\right) - \left(\frac{du}{d\theta}\right)_0\right\}\ln\coth\frac{|\theta|}{2}\,d\theta \qquad [393]$$

Here the first term represents the major contribution to the function $v(y)$. If $u(\theta)$ is a symmetrical function about a given point $\eta = y$ (that is, $\theta = 0$), then $(du/d\theta) - (du/d\theta)_0$ is an odd function about this point and (because $\ln\coth[|\theta|/2]$ is even about $\theta = 0$) the contribution of the integral in Eq. 393 becomes zero. Its value in any case is seen to depend

*Making the change of variable $x = \coth\theta/2$, one finds

$$2\int_0^{\infty}\ln\coth\frac{\theta}{2}\,d\theta = -4\int_1^{\infty}\frac{\ln x\,dx}{1-x^2} = \frac{\pi^2}{2}$$

which is No. 3 of Table 187 in Bierens de Haan, *Nouvelles tables d'intégrales définies* (Engels, Leiden), 1867.

only upon the asymmetry of the function $u(\theta)$ about the point $\theta = 0$, and because of the critical character of the factor ln coth $(|\theta|/2)$, only the asymmetry in the more immediate vicinity of this point can have an appreciable effect.

Although the computational labor involved in the solution of many problems of this sort may be quite heavy, the necessary theoretical basis for such a solution is evidently afforded by the Hilbert transforms or by modifications of these. However, when certain aspects of electric circuit theory are dealt with, a somewhat more complicated version of the problem presents itself. Instead of either $u(y)$ or $v(y)$ being specified over the entire range $0 < y < \infty$, $u(y)$ is specified only over portions of this range and $v(y)$ is specified over the remainder. For example, $u(y)$ may be specified for $0 < y < y_0$ and $v(y)$ for $y_0 < y < \infty$. The problem is to determine $u(y)$ and $v(y)$ for the ranges in which they are not specified,* that is, the object again is to determine the whole complex function over the entire range $0 < y < \infty$.

The clue to the method of solution lies in dividing the complex function $u(y) + jv(y)$ by the irrational factor $\sqrt{1 - y^2/y_0^2}$, which is real over the range $0 < y < y_0$ and imaginary over the complementary range $y_0 < y < \infty$. Hence the resulting function reads

$$\frac{u(y)}{\sqrt{1 - \dfrac{y^2}{y_0^2}}} + j\frac{v(y)}{\sqrt{1 - \dfrac{y^2}{y_0^2}}} \quad \text{for } 0 < y < y_0 \qquad [394]$$

and

$$\frac{v(y)}{\sqrt{\dfrac{y^2}{y_0^2} - 1}} - j\frac{u(y)}{\sqrt{\dfrac{y^2}{y_0^2} - 1}} \quad \text{for } y_0 < y < \infty \qquad [395]$$

The significant part about this result is that the real and imaginary parts of the resulting function alternately involve u and v in the two ranges. The same effect may, of course, be obtained if the complex function $u + jv$ is multiplied instead of divided by the irrational factor, but the resulting real and imaginary parts might then no longer remain finite in the vicinity of the point $y = \infty$. Also, one could presumably find other irrational functions which are alternately real and imaginary in the ranges $0 < y < y_0$ and $y_0 < y < \infty$, but the one used above is perhaps the simplest.

Substituting the real and imaginary parts of the functions 394 and

*The solution to this variation of the problem was obtained by H. W. Bode, *loc. cit.*

395 for $u(y)$ and $v(y)$ in the transforms 369 and 370, one finds respectively

$$-\frac{2}{\pi}\int_0^{y_0}\frac{\eta v(\eta)\,d\eta}{(\eta^2-y^2)\sqrt{1-\dfrac{\eta^2}{y_0^2}}}+\frac{2}{\pi}\int_{y_0}^{\infty}\frac{\eta u(\eta)\,d\eta}{(\eta^2-y^2)\sqrt{\dfrac{\eta^2}{y_0^2}-1}}=\begin{cases}\dfrac{u(y)}{\sqrt{1-\dfrac{y^2}{y_0^2}}} & \text{for } 0<y<y_0\\[2ex] \dfrac{v(y)}{\sqrt{\dfrac{y^2}{y_0^2}-1}} & \text{for } y_0<y<\infty\end{cases}\qquad [396]$$

and

$$\frac{2y}{\pi}\int_0^{y_0}\frac{u(\eta)\,d\eta}{(\eta^2-y^2)\sqrt{1-\dfrac{\eta^2}{y_0^2}}}+\frac{2y}{\pi}\int_{y_0}^{\infty}\frac{v(\eta)\,d\eta}{(\eta^2-y^2)\sqrt{\dfrac{\eta^2}{y_0^2}-1}}=\begin{cases}-\dfrac{v(y)}{\sqrt{1-\dfrac{y^2}{y_0^2}}} & \text{for } 0<y<y_0\\[2ex] \dfrac{-u(y)}{\sqrt{\dfrac{y^2}{y_0^2}-1}} & \text{for } y_0<y<\infty\end{cases}\qquad [397]$$

Equation 397 is used when $u(y)$ is specified over the range $0 < y < y_0$ and $v(y)$ over the range $y_0 < y < \infty$; Eq. 396 applies when the reverse is true. As an illustrative example, one may suppose that $u(y)$ is specified to be constant throughout the range $0 < y < y_0$ and that $v(y)$ is constant over the complementary range. Inasmuch as $u(y)$ is, by the present methods, determined within an arbitrary constant, one may simplify these given data somewhat by assuming $u(y)$ to be zero over the range $0 < y < y_0$ and adding the desired constant value to the whole solution afterward. Thus one has for the given data

$$\begin{aligned} u(y) &= 0 \quad \text{for } 0 < y < y_0 \\ v(y) &= v_0 \quad \text{for } y_0 < y < \infty \end{aligned} \qquad [398]$$

When Eq. 397 is applied, the first integral drops out because $u = 0$. The remaining expressions are somewhat simplified by the substitutions

$$x = \frac{\eta}{y_0} \qquad x_0 = \frac{y}{y_0} \qquad [399]$$

One then has

$$\frac{2v_0x_0}{\pi}\int_1^\infty \frac{dx}{(x^2 - x_0^2)\sqrt{x^2 - 1}} = \begin{cases} \dfrac{v(x_0)}{\sqrt{1 - x_0^2}} & \text{for } 0 < x_0 < 1 \\[2ex] \dfrac{-u(x_0)}{\sqrt{x_0^2 - 1}} & \text{for } 1 < x_0 < \infty \end{cases} \qquad [400]$$

The integration may be accomplished through noting first that

$$x_0\int_1^\infty \frac{dx}{(x^2 - x_0^2)\sqrt{x^2 - 1}} = \frac{1}{2}\int_1^\infty \frac{dx}{(x - x_0)\sqrt{x^2 - 1}} - \frac{1}{2}\int_1^\infty \frac{dx}{(x + x_0)\sqrt{x^2 - 1}} \qquad [401]$$

Using the substitution

$$z = \frac{xx_0 - 1}{x - x_0} \qquad [402]$$

one finds

$$\frac{dz}{\sqrt{1 - z^2}} = \frac{\sqrt{1 - x_0^2}\,dx}{(x - x_0)\sqrt{x^2 - 1}} \qquad [403]$$

whereas with

$$z = \frac{xx_0 + 1}{x + x_0} \qquad [404]$$

there results

$$\frac{dz}{\sqrt{1 - z^2}} = \frac{-\sqrt{1 - x_0^2}\,dx}{(x + x_0)\sqrt{x^2 - 1}} \qquad [405]$$

The changes in the limits of integration which accompany the substitutions 402 and 404 being noted, the integral on the left-hand side of Eq. 401 is found to be given by

$$\frac{1}{2\sqrt{1 - x_0^2}}\left[\int_{-1}^{x_0} \frac{dz}{\sqrt{1 - z^2}} + \int_1^{x_0} \frac{dz}{\sqrt{1 - z^2}}\right] \qquad [406]$$

The integrands in these integrals are even functions. Hence the integral from 1 to x_0 may be replaced by one from $-x_0$ to -1. The two integrals in the expression 406 may then be combined, giving

$$\frac{2v_0x_0}{\pi}\int_1^\infty \frac{dx}{(x^2 - x_0^2)\sqrt{x^2 - 1}} = \frac{v_0}{\pi\sqrt{1 - x_0^2}}\int_{-x_0}^{x_0} \frac{dz}{\sqrt{1 - z^2}} \qquad [407]$$

The integration is now no longer difficult. Equation 400 is thus observed to yield for $0 < x_0 < 1$

$$v(x_0) = \frac{2v_0}{\pi} \sin^{-1} x_0 = \frac{2v_0}{\pi} \ln \left(\sqrt{1 - x_0^2} + jx_0\right) \qquad [408]$$

whereas for $1 < x_0 < \infty$, in which range it is appropriate to consider the right-hand side of Eq. 407 in the form

$$\frac{-v_0}{\pi\sqrt{x_0^2 - 1}} \int_{-x_0}^{x_0} \frac{dz}{\sqrt{z^2 - 1}} \qquad [409]$$

the integration yields the real part

$$u(x_0) = \frac{2v_0}{\pi} \ln \left(\sqrt{x_0^2 - 1} + x_0\right) \qquad [410]$$

One readily recognizes from these results that the complete complex function is given over the entire range $0 < y < \infty$ by the expression

$$u(y) + jv(y) = \frac{2v_0}{\pi} \ln \left(\sqrt{1 - \frac{y^2}{y_0^2}} + \frac{jy}{y_0}\right) \qquad [411]$$

The procedure may readily be extended to problems in which the range $0 < y < \infty$ is divided into more than two subranges. For example, if u and v are alternately specified in the ranges $0 < y < y_1$, $y_1 < y < y_2$, $y_2 < y < \infty$, the Hilbert transforms are applied to the real and imaginary parts of the function $(u + jv)$ divided by $\sqrt{(1 - y^2/y_1^2)(1 - y^2/y_2^2)}$. The determination of the appropriate detailed relationships follows the same pattern as given above for the simpler problem involving two subranges.

23. More about potential theory and conjugate functions

Additional useful physical interpretations relevant to properties of functions of a complex variable may be had from pursuing further the intimate relationships between certain of these properties and potential theory. The starting point for the present discussion is the simple physical situation shown in Fig. 23. The plane of the paper, which is also regarded as the complex z-plane, represents a plane perpendicular to an infinitely long straight conductor. The origin of co-ordinates for this plane is chosen coincident with the conductor, which is assumed to carry a uniform linear distribution of charge of q coulombs per unit length and also a uniform current of i amperes whose reference direction points upward.

According to elementary theory, the electric field intensity $\mathfrak{E}$ and the magnetic flux density $\mathfrak{B}$ at the point z_0 are given by the negative gradient

of corresponding potential functions $\mathcal{V}$ and $\mathcal{F}$, that is,

$$\mathcal{E} = -\operatorname{grad} \mathcal{V} \qquad \mathcal{B} = -\operatorname{grad} \mathcal{F} \tag{412}$$

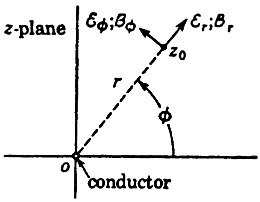


Fig. 23. The plane perpendicular to an infinitely long straight charged conductor carrying current.

with

$$\mathcal{V} = \frac{2q}{\varepsilon} \ln \frac{1}{r} \tag{413}$$

and

$$\mathcal{F} = -2\mu i\phi \tag{414}$$

in which ε and μ are the dielectric permittivity and magnetic permeability respectively of the homogeneous isotropic surrounding medium.* The field components in polar co-ordinates are

$$\mathcal{E}_r = -\frac{\partial \mathcal{V}}{\partial r}; \qquad \mathcal{E}_\phi = -\frac{\partial \mathcal{V}}{r\,\partial \phi} \tag{415}$$

$$\mathcal{B}_r = -\frac{\partial \mathcal{F}}{\partial r}; \qquad \mathcal{B}_\phi = -\frac{\partial \mathcal{F}}{r\,\partial \phi} \tag{416}$$

By substitution from Eqs. 413 and 414 these become

$$\mathcal{E}_r = \frac{2q}{\varepsilon r}; \qquad \mathcal{E}_\phi = 0 \tag{417}$$

$$\mathcal{B}_r = 0; \qquad \mathcal{B}_\phi = \frac{2\mu i}{r} \tag{418}$$

It is now useful to observe that if one defines a single complex potential function

$$w = \frac{\varepsilon \mathcal{V}}{2q} + j\frac{\mathcal{F}}{2\mu i} = -\ln re^{j\phi} = \ln \frac{1}{z_0} \tag{419}$$

in which

$$z_0 = re^{j\phi} \tag{420}$$

the field components are determined from

$$-\frac{\partial w}{\partial r} = \frac{\varepsilon \mathcal{E}_r}{2q} + j\frac{\mathcal{B}_r}{2\mu i} \tag{421}$$

and

$$-\frac{\partial w}{r\,\partial \phi} = \frac{\varepsilon \mathcal{E}_\phi}{2q} + j\frac{\mathcal{B}_\phi}{2\mu i} \tag{422}$$

*Unrationalized MKS units are used in these discussions.

or the resultant fields from the relation

$$-\text{grad } w = \frac{\varepsilon\mathcal{E}}{2q} + j\frac{\mathcal{B}}{2\mu i} \qquad [423]$$

The real and imaginary parts of the gradient of the complex potential function w yield the electric and magnetic fields respectively (except for constant multipliers).

It is thus seen that the real and imaginary parts u and v of the simple function

$$w = \ln\frac{1}{z_0} = u + jv \qquad [424]$$

when multiplied respectively by the constant factors $2q/\varepsilon$ and $2\mu i$ are the scalar electric and magnetic potential functions for the single charge and current-carrying conductor of Fig. 23. This example, incidentally, affords an interesting physical interpretation of the multivalued character of the logarithm function. Since, for the scalar magnetic potential $\mathcal{F}$, the radial lines are equipotential loci, this function changes continuously in the same direction as one proceeds along a path encircling the conductor, each complete revolution yielding an increment proportional to 2π. If $\mathcal{F}$ is likened to an altitude function in a physical terrain, the similarity of the Riemann surface to a winding ramp is realized. The branch point about which the surface winds becomes, in terms of this magnetic field interpretation, a vortex thread which is physically realized by the current-carrying conductor. This singularity for the field and the function representing it is indeed the seat of its excitation.

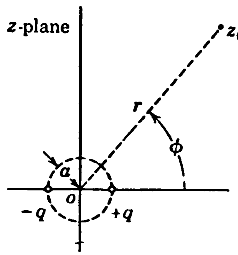


FIG. 24. The plane perpendicular to a pair of identical infinitely long conductors carrying current in opposite directions and carrying opposite charges.

A pair of identical conductors are now considered. They are spaced symmetrically with respect to the origin as shown in Fig. 24 and are assumed to carry charges and currents of equal magnitude but opposite sign. The complex potential function in this case is, with reference to the notation indicated in the figure,

$$w = \ln\frac{1}{z_0 - a} - \ln\frac{1}{z_0 + a} \qquad [425]$$

or

$$w = \ln\left(\frac{z_0 + a}{z_0 - a}\right) \qquad [426]$$

It is now assumed that the spacing $2a$ between the conductors is very small compared with the distance to any point z_0 at which the field (or the resulting function) is to be studied. For the moment this spacing may be regarded as being finite, although for some of the subsequent discussions it is more appropriate to allow the limiting process $a \to 0$ to be completed. At all events the condition

$$\frac{a}{|z_0|} \ll 1 \qquad [427]$$

is assumed to be fulfilled, so that Eq. 426 is replaceable by

$$w \cong \ln\left(1 + \frac{2a}{z_0}\right) \cong \frac{2a}{z_0} \qquad [428]$$

with any accuracy that may become necessary. This form for w may be obtained through expanding the logarithm in a Maclaurin series in terms of the variable $2a/z_0$ and retaining only the first term, or it may be derived in a somewhat more lucid fashion through first writing the right-hand side of Eq. 425 in terms of the defining integral for the logarithm, thus

$$w = \ln(z_0 + a) - \ln(z_0 - a) = \int_1^{z_0+a} \frac{d\zeta}{\zeta} - \int_1^{z_0-a} \frac{d\zeta}{\zeta} \qquad [429]$$

which is equivalent to

$$w = \int_{z_0-a}^{z_0+a} \frac{d\zeta}{\zeta} \qquad [430]$$

If the condition 427 is fulfilled, then the points $z_0 + a$ and $z_0 - a$ are so close together that this integral is very nearly given by

$$w = \frac{1}{z_0}\int_{z_0-a}^{z_0+a} d\zeta = \frac{2a}{z_0} \qquad [431]$$

which agrees with Eq. 428.

As stated above, the conductors in Fig. 24 also carry currents i having equal magnitudes and opposite directions (upward from the plane of the paper for the conductor carrying $+q$) so that the real and imaginary parts of w again represent scalar electric and magnetic potential functions when multiplied by their appropriate factors. These, incidentally, may just as well be assumed to be numerically equal, as expressed by

$$\frac{q}{i} = \varepsilon\mu \qquad \text{(numerically)} \qquad [432]$$

Then the common factor may be included in the expression for w. Since this procedure facilitates matters somewhat, it is assumed for the follow-

ing discussion that

$$w = \frac{4qa}{\varepsilon z_0} \qquad [433]$$

It is now possible to let the spacing $2a$ become as small as desired without having the magnitude of w become small at the same time, since one can demand that as $2a$ is made smaller and smaller, the charge q be made correspondingly larger so that the product

$$2aq = m \qquad [434]$$

remains constant. In the limit $a \to 0$, q approaches infinity. This limiting configuration of oppositely charged conductors is a useful concept and is referred to as a ***dipole*** or ***doublet***. The quantity m, called the ***moment*** of the dipole, is considered to be a vector coincident with the line joining the two charges (the *axis* of the dipole) and directed from the negative to the positive charge. The vector character of this dipole moment is taken into account through its being considered a complex number although in the present example it is real, as is evident from inspection of Fig. 24.

The complex potential function for the dipole is

$$w = \frac{2m}{\varepsilon z_0} = \frac{2m}{\varepsilon r} e^{-j\phi} \qquad [435]$$

and the scalar electric and magnetic potential functions are the real and imaginary parts

$$u = \frac{2m}{\varepsilon r} \cos \phi \qquad [436]$$

and

$$v = -\frac{2m}{\varepsilon r} \sin \phi \qquad [437]$$

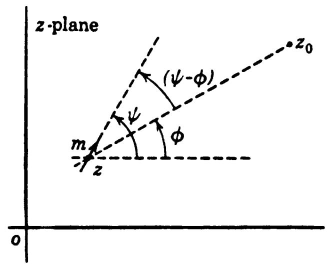


FIG. 25. A dipole of moment m with arbitrary position and orientation.

Although in this derivation the dipole is assumed to have a very particular location and orientation in the z-plane, namely, that shown in Fig. 24, one may readily generalize the expression 435 so that it becomes appropriate to a dipole in an arbitrary position and orientation as is shown in Fig. 25, where the dipole moment makes an angle ψ with the reference axis. The dipole is located at an arbitrary point z, so that one now has

$$(z_0 - z) = re^{j\phi} \qquad [438]$$

With m denoting the magnitude of the dipole moment, it is readily seen

that the appropriate modification of Eq. 435 reads

$$w = \frac{2me^{j\psi}}{\varepsilon(z_0 - z)} = \frac{2m}{\varepsilon r} e^{j(\psi-\phi)} \tag{439}$$

which has the real and imaginary parts

$$u = \frac{2m}{\varepsilon r} \cos(\psi - \phi) \tag{440}$$

and

$$v = \frac{2m}{\varepsilon r} \sin(\psi - \phi) \tag{441}$$

It is significant that if the dipole is rotated through 90 degrees in the clockwise direction — that is, if ψ is replaced by $\psi - (\pi/2)$ — then u becomes replaced by v, and v by $-u$. In other words, rotation of the dipole through 90 degrees in the clockwise direction has the effect of replacing each potential function by its conjugate. Further discussion of this useful observation is given presently.

The preceding derivations also show that the dipole may be regarded as the result of merging two logarithmic singularities of opposite polarity and infinite magnitude. As Eq. 439 shows, the end result is a simple pole, and the complex residue of the function w in this pole, except for the factor $2/\varepsilon$ is numerically equal to the vector moment $me^{j\psi}$ of the dipole. One thus obtains an interesting physical interpretation for the kind of singularity designated as a simple pole, namely, that it may be visualized as the seat of an electric or magnetic dipole with its moment proportional to the residue of the function in this pole.

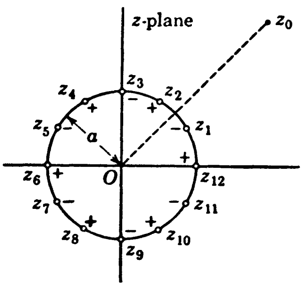


FIG. 26. A $2n$-pole centered at the origin. The position of + and − charges is correlated to the form of the resultant complex potential function.

It is logical to continue with a similar study of poles of a higher order of multiplicity. As the following discussion shows, these are represented by the merging of symmetrical configurations of larger numbers of charges having equal (infinite) magnitudes but signs which are alternately positive and negative, there being an equal number of each. For example, the configuration following the dipole in its order of complexity is the so-called quadripole, consisting of two positive and two negative charges arranged alternately and spaced symmetrically on the circumference of a (vanishingly) small circle of radius a. Next in the order of complexity is the configuration involving six charges of which three are

positive and three negative, and so on. In general such a configuration is referred to as a $2n$-pole, n being the number of positive or negative charges. The dipole is a $2n$-pole for $n = 1$; the quadripole is a $2n$-pole for $n = 2$, and so forth.

Figure 26 shows a $2n$-pole for $n = 6$ with its center at the origin of the z-plane. The points $z_1, z_2, \cdots z_{2n}$ are uniformly spaced on the circumference of the small circle, the even-numbered ones corresponding to positive charges and the odd-numbered ones to negative charges. The resultant complex potential function is given by

$$w = \frac{2q}{\varepsilon}\left[\ln\frac{1}{z_0 - z_2} + \ln\frac{1}{z_0 - z_4} + \cdots + \ln\frac{1}{z_0 - z_{2n}} - \ln\frac{1}{z_0 - z_1} - \ln\frac{1}{z_0 - z_3} - \cdots - \ln\frac{1}{z_0 - z_{2n-1}}\right] \quad [442]$$

or

$$w = \frac{2q}{\varepsilon}\ln\frac{(z_0 - z_1)(z_0 - z_3)\cdots(z_0 - z_{2n-1})}{(z_0 - z_2)(z_0 - z_4)\cdots(z_0 - z_{2n})} \quad [443]$$

Now it is recognized that the values of $z_1, z_3, \cdots z_{2n-1}$ are the roots of the equation

$$z^n + a^n = 0 \quad [444]$$

which are

$$z = a\sqrt[n]{-1} = ae^{j(\nu\pi/n)} \qquad (\nu = 1, 3, 5, \cdots 2n - 1) \quad [445]$$

and the values of $z_2, z_4, \cdots z_{2n}$ are the roots of the equation

$$z^n - a^n = 0 \quad [446]$$

which are

$$z = a\sqrt[n]{1} = ae^{j(\nu\pi/n)} \qquad (\nu = 2, 4, 6, \cdots 2n) \quad [447]$$

Consequently,

$$(z_0 - z_1)(z_0 - z_3)\cdots(z_0 - z_{2n-1}) = z_0^n + a^n \quad [448]$$

and

$$(z_0 - z_2)(z_0 - z_4)\cdots(z_0 - z_{2n}) = z_0^n - a^n \quad [449]$$

so that Eq. 443 becomes

$$w = \frac{2q}{\varepsilon}\ln\left(\frac{z_0^n + a^n}{z_0^n - a^n}\right) \quad [450]$$

Again if

$$\frac{a}{|z_0|} \ll 1 \quad [451]$$

the logarithm in this expression is very nearly given by

$$w = \frac{2q}{\varepsilon} \ln\left(1 + \frac{2a^n}{z_0{}^n}\right) \cong \frac{4qa^n}{\varepsilon z_0{}^n} \qquad [452]$$

If the moment of the $2n$-pole is defined as

$$m = 2qa^n \qquad [453]$$

its complex potential function becomes

$$w = \frac{2m}{\varepsilon z_0{}^n} = \frac{2m}{\varepsilon r^n} e^{-jn\phi} \qquad [454]$$

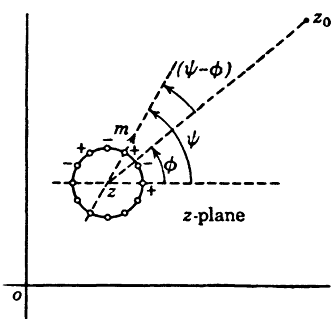


FIG. 27. A $2n$-pole centered at z.

which yields the result expressed by Eq. 435 for $n = 1$.

When the $2n$-pole is located at an arbitrary point z and has an arbitrary orientation, as shown in Fig. 27, its complex potential at the point z_0 is given by

$$w = \frac{2me^{jn\psi}}{\varepsilon(z_0 - z)^n} \qquad [455]$$

because z_0 in Eq. 452 becomes replaced by $(z_0 - z)e^{-j\psi}$. Here the angle ψ is evidently determined only within an additive positive or negative integer multiple of the angle $2\pi/n$.

If again one writes

$$(z_0 - z) = re^{j\phi} \qquad [456]$$

when

$$w = \frac{2m}{\varepsilon r^n} e^{jn(\phi-\psi)} \qquad [457]$$

so that

$$u = \frac{2m}{\varepsilon r^n} \cos n(\psi - \phi) \qquad [458]$$

and

$$v = \frac{2m}{\varepsilon r^n} \sin n(\psi - \phi) \qquad [459]$$

According to Eq. 455, the $2n$-pole evidently represents a pole of multiplicity n for the function w.

It is now of interest to consider a configuration of charges resulting when many identical dipoles are placed side by side, with a uniform infinitesimal spacing, their centers lying on any curve and their moments all pointing toward the same side of this curve. Such an arrangement is shown in Fig. 28. If the dimension perpendicular to the plane of the paper, in which longitudinal uniformity obtains, is mentally supplied to this picture, it is readily recognized that this arrangement may be regarded as

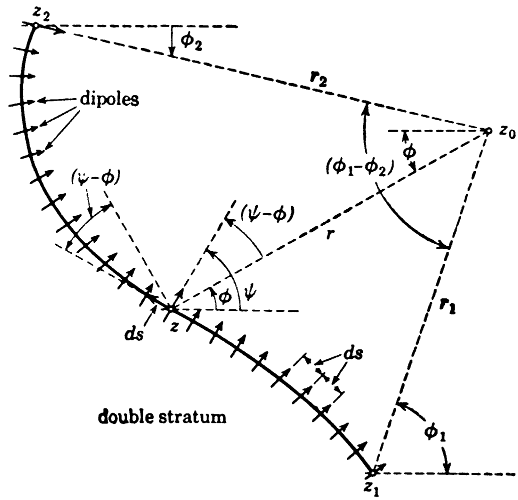


FIG. 28. The cross section of a sheet having a uniform surface distribution of dipoles.

a uniform surface distribution of dipoles, and that the result is equivalent to having a surface or membrane uniformly covered with a charge of equal magnitude but opposite sign on its two sides, the thickness of the membrane corresponding to the spacing $2a$ between the charges in the dipoles.

This configuration is called a uniform *double stratum*. If the magnitude of the surface density of charge is denoted by σ, and $2a$ is the thickness of the membrane, or the separation of the surface charges of opposite polarity,

$$2a\sigma = \tau \qquad [460]$$

is defined as the moment of the double stratum. In terms of the individual dipoles with their charges $\pm q$, one recognizes that if the uniform spacing between the dipoles is denoted by the differential ds,

$$q = \sigma\, ds \qquad [461]$$

and hence

$$2aq = m = \tau\, ds \qquad [462]$$

The problem is to determine the scalar electric potential u at the point z_0. This may be determined through integrating the expression 440 over the surface of the double stratum. Thus, the differential contribution to the net potential at z_0 due to the single dipole at the point z is, with the help of Eq. 462,

$$du = \frac{2m}{\varepsilon r}\cos(\psi - \phi) = \frac{2\tau\, ds \cos(\psi - \phi)}{\varepsilon r} \qquad [463]$$

But from Fig. 28 it is recognized that

$$-r\, d\phi = ds \cos(\psi - \phi) \qquad [464]$$

so that Eq. 463 becomes

$$du = -\frac{2\tau}{\varepsilon}\, d\phi \qquad [465]$$

Hence one obtains for the net potential at the point z_0

$$u = \int_{z_1}^{z_2} du = -\frac{2\tau}{\varepsilon}\int_{\phi_1}^{\phi_2} d\phi = \frac{2\tau}{\varepsilon}(\phi_1 - \phi_2) \qquad [466]$$

This remarkably simple result states that the potential is proportional only to the net angle subtended at the point z_0 by the boundaries of the double stratum. The shape of the double stratum between these boundaries is immaterial. It is essential to the achievement of this simple result that the moment τ of the double stratum be constant over its entire surface. The expression given by Eq. 466 applies only to the potential of a *uniform* double stratum.

It is now a simple matter to find the potential function which is the conjugate of this potential u of the double stratum. It is pointed out above in connection with the single dipole that the conjugate to the electric potential function u results when that dipole is rotated through 90 degrees in the clockwise direction. According to the principle of superposition, the conjugate to the double-stratum potential results when all the dipoles constituting it are rotated through 90 degrees. This process converts the double stratum into a *dipole chain.* If one chooses to make the spacing ds between the dipoles in the double stratum equal to the spacing $2a$ between the charges in the individual dipoles, that is, if

$$ds = 2a \qquad [467]$$

the adjacent positive and negative charges in the dipole chain fall on top of one another and hence cancel, so that there is left only one negative charge $-q$ at the point z_2 and one positive charge q at the point z_1.

The choice indicated by Eq. 467 is, according to Eq. 462, equivalent to making

$$q = \tau \tag{468}$$

so that the potential function conjugate to u given by Eq. 466 becomes

$$v = \frac{2\tau}{\varepsilon} \ln \frac{r_2}{r_1} \tag{469}$$

Since replacing u by v and v by $-u$ again yields a pair of conjugate potential functions, one may also write

$$u = \frac{2\tau}{\varepsilon} \ln \frac{r_1}{r_2} \tag{470}$$

and

$$v = \frac{2\tau}{\varepsilon} (\phi_1 - \phi_2) \tag{471}$$

These are evidently the real and imaginary parts of the complex function

$$w = \frac{2\tau}{\varepsilon} \ln \left(\frac{z - z_1}{z - z_2} \right) \tag{472}$$

in which z_0 of Fig. 28 is now considered to be any variable point z.

The response functions of linear passive electrical networks (driving point and transfer impedances or admittances) are rational functions of the form

$$f(z) = \frac{(z - z_1)(z - z_3) \cdots (z - z_{2m-1})}{(z - z_2)(z - z_4) \cdots (z - z_{2m})} \tag{473}$$

In problems of this sort it is expedient to write

$$f(z) = e^{\theta} = e^{A+jB} \tag{474}$$

and have, for the so-called *loss* and *angle* functions A and B,

$$= A + jB = \ln \frac{(z - z_1)(z - z_3) \cdots (z - z_{2m-1})}{(z - z_2)(z - z_4) \cdots (z - z_{2m})} \tag{475}$$

Thus the loss function A may be regarded as the electric potential resulting when a system of identical charged conductors are placed at the points $z_2, z_4, \cdots z_{2m}$ and a set of conductors with charges of opposite sign are placed at the points $z_1, z_3, \cdots z_{2m-1}$. The function A is usually studied for pure imaginary values of z, that is, for points on the imaginary axis in the z-plane (these correspond to real frequencies in steady state circuit analysis). B is then the conjugate potential function with respect to A. A problem in network design is thus reduced to the determination of a distribution of the points $z_1, z_2, z_3, \cdots$ which yields suitable functions

A and B. The visualization of the problem lent by the analogy to potential theory can in many cases be helpful.

If the rational function, Eq. 473, is expanded in partial fractions (as shown in Art. 16), each term is recognized as having the form of Eq. 439. The real and imaginary parts of this function may, therefore, be interpreted as the potentials due to dipoles at the points $z_2, z_4, \cdots z_{2m}$. More specifically, the real part (which may be a resistance or conductance) is the potential of the dipole distribution represented by the partial fraction expansion of $f(z)$, whereas the imaginary part (a corresponding reactance or susceptance) is the conjugate potential resulting after all the dipoles are rotated through 90 degrees in the clockwise direction.

24. Some useful functions in conformal mapping; the linear fractional function

The artifice of introducing a change of independent variable by substituting for the given one some function of a new variable usually implies the transformation of one region of the complex plane into another. A contour in the one region is thereby mapped upon the other. The two regions may be mutually exclusive, or partly or wholly overlapping, depending upon the nature of the transforming function. A very common substitution is

$$\zeta = \frac{1}{z} \qquad [476]$$

which is more explicitly written

$$\zeta = \xi + j\eta = \frac{1}{x + jy} = \frac{x - jy}{x^2 + y^2} \qquad [477]$$

whence

$$\xi = \frac{x}{|z|^2}, \qquad \eta = \frac{-y}{|z|^2} \qquad [478]$$

In polar co-ordinates

$$z = re^{j\phi} \qquad [479]$$

$$\zeta = \rho e^{j\theta} \qquad [480]$$

so that the transformation 476 yields

$$\rho = \frac{1}{r} \qquad [481]$$

and

$$\theta = -\phi \tag{482}$$

Because of the relation 481, this substitution is spoken of as a transformation in terms of *reciprocal radii*. Since ρ is larger than, equal to, or smaller than unity as r is respectively smaller than, equal to, or larger than unity, it is seen that the region of the z-plane within the unit circle is transformed into the region outside this circle, and vice versa, whereas the unit circle itself remains intact except for a shifting of specific points according to the relation 482. In particular, the points ± 1 on the real axis remain fixed by the transformation, and these are consequently spoken of as the *fixed points*.

A significant property of the transformation 476 is that the origin and the point at infinity become interchanged, that is, $z = 0$ corresponds to $\zeta = \infty$, and $z = \infty$ corresponds to $\zeta = 0$. Thus a given function $f(z)$ may be studied in the vicinity of $z = \infty$ through introducing the change of variable 476 and then studying the resulting function $F(\zeta) = f(1/z)$ in the vicinity of $\zeta = 0$.

A variation of this transformation reads

$$\zeta^* = \frac{1}{\bar{z}} \tag{483}$$

in which the bar indicates the conjugate value. With

$$\zeta^* = \rho^* e^{j\theta^*} \tag{484}$$

this yields

$$\rho^* = \frac{1}{r} \tag{485}$$

and

$$\theta^* = \phi \tag{486}$$

All the points on the unit circle are now fixed points. Corresponding points z and ζ^* may be found by means of the geometrical construction indicated in Fig. 29, as may be seen from the fact that the similar right-angle triangles yield the relation

$$\frac{|z|}{OA} = \frac{OA}{|\zeta^*|} \tag{487}$$

in which OA is the radius of the unit circle and hence equal to unity. The point ζ^* is said to be the *image* of the point z with respect to the unit circle. The transformation indicated by Eq. 483 is, therefore, also spoken of as a *reflection* or *inversion* with respect to the unit circle. As the point z traces

some figure inside the unit circle, the corresponding point ζ^* maps the image of this figure outside the unit circle.

If the unit circle is replaced by a circle of radius R, the transformation reads $\zeta^* = R^2/\bar{z}$. The geometrical construction shown in Fig. 29 applies to any value of R. If $R - |z|$ is held constant as R is allowed to become larger and larger, and attention is focused upon that part of Fig. 29 in the vicinity of the points z and ζ^* alone, the circle will appear to degenerate into a straight line and the two points in question will ultimately be symmetrically located with respect to this line in the sense of an object and its mirror image.

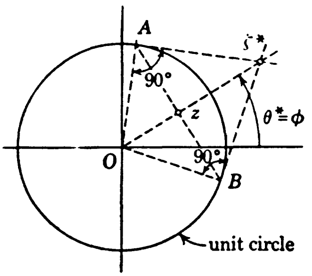


FIG. 29. The point z and its image $\zeta^* = 1/\bar{z}$ as obtained by graphical construction.

Other variations of this transformation which are sometimes used are $\zeta' = -\zeta$ and $\zeta^{*\prime} = -\zeta^*$. All these substitutions have the property of transforming circles into circles, as is shown from a more general point of view in the subsequent discussion.

The transformations $\zeta = \pm 1/z$ are, of course, conformal, and hence, as shown in Art. 2, angular relationships are preserved in sense as well as in magnitude. The transformations $\zeta^* = \pm 1/\bar{z}$ are isogonal but not conformal. They preserve angular relationships except for a reversal of the positive sense of angular measurement, as in the right- and left-hand relationship between an object and its mirror image. For any closed contour not encircling the origin,* the direction of traversal is preserved by the transformations $\zeta = \pm 1/z$, whereas it is reversed in the transformations $\zeta^* = \pm 1/\bar{z}$. These relationships are indicated for a set of circles in Fig. 30, in which a typical point P on the given locus z is also indicated in each of the transformed loci.

It is collaterally useful to observe that, for the transformations $\zeta' = -1/z$ and $\zeta^* = 1/\bar{z}$, a given locus lying wholly in the upper (respectively lower) half plane, remains wholly within that half plane. In these cases, the upper (respectively lower) half plane is said to be transformed into itself. Similarly, the substitution $\zeta = 1/z$ transforms the right (respectively left) half plane into itself, whereas $\zeta^{*\prime} = -1/\bar{z}$ transforms the right into the left half plane, and vice versa. In a specific problem the particular form of the substitution chosen is thus seen to depend upon the detailed requirements set by the nature of that problem and the purpose for which the substitution or transformation is made.

*The conclusions in this statement are reversed if the contour does enclose the origin. This fact may readily be appreciated from the study of a few simple examples.

The substitutions $\zeta = \pm 1/z$ are special cases of the more general *linear fractional* or *bilinear* form

$$\zeta = \frac{az + b}{cz + d} \tag{488}$$

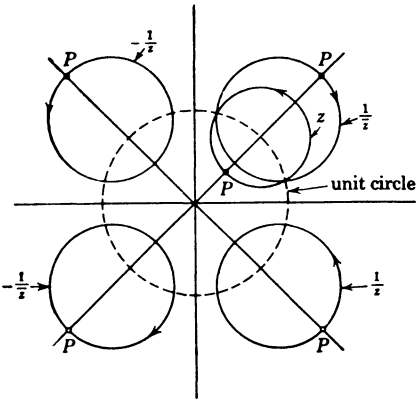


FIG. 30. Corresponding circular loci of z and various related reciprocals. Note the various senses of angular relationships.

in which the complex constants a, b, c, d are subject to the condition that the so-called determinant $ad - bc$ of the transformation shall not vanish, that is,

$$ad - bc \neq 0 \tag{489}$$

and, in addition,

$$c \neq 0 \tag{490}$$

The function 488 then is regular everywhere except at the one point

$$z = -\frac{d}{c} \tag{491}$$

at which it has a simple pole.

Inclusive of this point, there is a unique correspondence between all points in the z- and ζ-planes, so that the inverse of the function 488 which reads

$$z = \frac{d\zeta - b}{-c\zeta + a} \tag{492}$$

has identical mapping properties.

The condition 490 is necessary in order to prevent 488 from degenerating into an ordinary linear form.* The reason for the condition 489 is appreciated through consideration of two points ζ_1 and ζ_2 in the ζ-plane and their corresponding points z_1 and z_2 in the z-plane. Equation 488 then yields

$$\zeta_1 - \zeta_2 = \frac{(ad - bc)(z_1 - z_2)}{(cz_1 + d)(cz_2 + d)} \tag{493}$$

whence it is clear that a unique correspondence of points does not result if $ad - bc = 0$.

In addition to ζ_1 and ζ_2, two more points ζ_3 and ζ_4 together with their corresponding points z_3 and z_4 are now considered. By analogy to the form of Eq. 493, it is then readily seen that

$$\frac{\zeta_3 - \zeta_1}{\zeta_3 - \zeta_2} = \frac{z_3 - z_1}{z_3 - z_2} \times \frac{cz_2 + d}{cz_1 + d} \tag{494}$$

and

$$\frac{\zeta_4 - \zeta_1}{\zeta_4 - \zeta_2} = \frac{z_4 - z_1}{z_4 - z_2} \times \frac{cz_2 + d}{cz_1 + d} \tag{495}$$

and hence that

$$\frac{\zeta_3 - \zeta_1}{\zeta_3 - \zeta_2} \bigg/ \frac{\zeta_4 - \zeta_1}{\zeta_4 - \zeta_2} = \frac{z_3 - z_1}{z_3 - z_2} \bigg/ \frac{z_4 - z_1}{z_4 - z_2} \tag{496}$$

With reference to Fig. 31 it is clear that

$$\frac{z_3 - z_1}{z_3 - z_2} = \frac{|z_3 - z_1|}{|z_3 - z_2|} e^{-j\alpha} \tag{497}$$

and

$$\frac{z_4 - z_1}{z_4 - z_2} = \frac{|z_4 - z_1|}{|z_4 - z_2|} e^{-j\beta} \tag{498}$$

Now if the points z_1, z_2, z_3, z_4 are assumed to lie on a circle, it follows that $\alpha = \beta$. The right-hand side of Eq. 496 is then seen to reduce to a positive real constant. Since, according to a familiar geometrical proposition, the angles α and β can be equal only if the four points in question lie on a circle, and since, according to Eq. 496, a similar geometrical relationship to that shown in Fig. 31 likewise applies to the quantities

*The mapping properties of the linear transformation $w = az + b$ are recognized as amounting to a magnification and rotation according to the magnitude and angle of the complex constant a, and a translation equal to the value of b. Any picture or map subjected to this transformation obviously remains undistorted. The mapping properties of the linear transformation are, therefore, said to be ***homographic***.

$\zeta_1, \zeta_2, \zeta_3, \zeta_4$, one may conclude that the corresponding points $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ must also lie on a circle. An important property of the transformation 488 or its inverse 492 is thus proved, namely, that this linear fractional form carries circles over into circles. In other words, a circular locus in the z-plane corresponds to a circular locus in the ζ-plane. Straight lines are included here since they represent circles with infinite radii, but it is not implied, of course, that straight lines are necessarily carried over into straight lines.

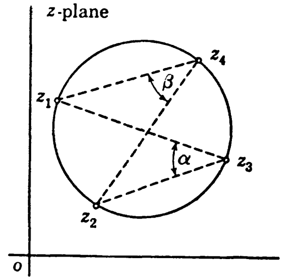


FIG. 31. Construction used in showing that the linear fractional transformation carries circles over into circles.

The linear fractional transformation is, incidentally, the most general analytic transformation yielding a one-to-one correspondence between all points in the simple w- and z-planes. The term " simple " w- or z-plane is intended to denote such a one in which the unique distinction between points does not require the concept of a multileaved Riemann surface. With the interpretative aid of such a surface, any analytic function yields a one-to-one correspondence between points in the w- and z-planes. In the simple complex plane however (" schlichte Ebene " in the German literature), a one-to-one correspondence between points is afforded only by the so-called *schlicht* functions, of which the linear fractional function is the most general form.

The *fixed points* of the transformation are those for which $\zeta = z$, or, Eq. 488 being used, they correspond to the z-values satisfying the equation

$$z = \frac{az + b}{cz + d} \tag{499}$$

which is equivalent to

$$cz^2 + (d - a)z - b = 0 \tag{500}$$

The roots of this quadratic equation may be denoted by z_1 and z_2. Then

$$\zeta_1 = z_1 \qquad \text{and} \qquad \zeta_2 = z_2 \tag{501}$$

If the ζ-plane is superimposed upon the z-plane, any circle passing through the points z_1 and z_2 is transformed into another circle passing through these same points. Hence the family of circles passing through the fixed points z_1 and z_2 is transformed into itself in the sense that any one of these circles is transformed into another belonging to the same family. Now these circles possess an orthogonal family of loci which are also circles. These enclose the points z_1 and z_2, as shown in Fig. 32. Since

the transformation is a conformal one, it follows that the transformation of an orthogonal circle must also be orthogonal to any of the circles through z_1 and z_2. Hence the family of orthogonal circles is transformed into itself also.

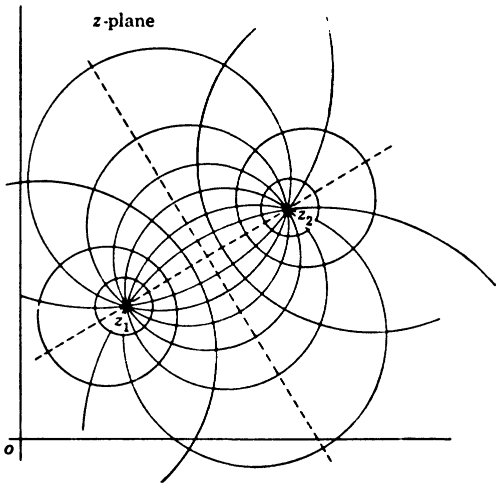


FIG. 32. Circles through the *fixed points* z_1 and z_2 are transformed into other circles through the same two *fixed points*.

For proper choices of the constants in Eq. 488, this transformation may be made to carry any specified circle in the z-plane over into an independently specified circle in the ζ-plane. In order to see this one may write by analogy to Eq. 496

$$\frac{\zeta - \zeta_1}{\zeta - \zeta_2} \Big/ \frac{\zeta_3 - \zeta_1}{\zeta_3 - \zeta_2} = \frac{z - z_1}{z - z_2} \Big/ \frac{z_3 - z_1}{z_3 - z_2} \tag{502}$$

in which z_1, z_2, z_3 are distinct and z is a variable point on a given circle in the z-plane, whereas ζ_1, ζ_2, ζ_3, and ζ are the corresponding distinct and variable points on the resulting circle in the ζ-plane. If one writes for the sake of abbreviation

$$H = \frac{\zeta_3 - \zeta_1}{\zeta_3 - \zeta_2} \Big/ \frac{z_3 - z_1}{z_3 - z_2} \tag{503}$$

Eq. 502, which is equivalent to the transformation 488, takes the form

$$\frac{\zeta - \zeta_1}{\zeta - \zeta_2} = H\frac{z - z_1}{z - z_2} \tag{504}$$

Here z_1, z_2, z_3 are any distinct points determining a specific circle in the z-plane, and ζ_1, ζ_2, ζ_3 are any chosen corresponding points in the ζ-plane determining a circle which the desired transformation is intended to yield.

In particular, the points z_1 and z_2 in Eq. 503 may be selected as the fixed points of the transformation. Then $\zeta_1 = z_1$ and $\zeta_2 = z_2$, so that

$$\mathrm{H} = \frac{\zeta_3 - z_1}{\zeta_3 - z_2} \bigg/ \frac{z_3 - z_1}{z_3 - z_2} \tag{505}$$

Equation 504 then reads

$$\frac{\zeta - z_1}{\zeta - z_2} = \mathrm{H}\,\frac{z - z_1}{z - z_2} \tag{506}$$

This is called the normal form of the transformation 488 or its inverse 492.

Two special types of this transformation may now be defined. For the first of them, any member of the family of circles passing through the fixed points z_1 and z_2 is transformed into itself. This statement means that the points z_3 and ζ_3 lie on the *same* circle, although they are, of course, not coincident like the points $\zeta_1 = z_1$ and $\zeta_2 = z_2$; otherwise the transformation would degenerate into the trivial identity $\zeta = z$. The points z_1, z_2, z_3, ζ_3 are then four distinct points on the same circle, like the four points z_1, z_2, z_3, z_4 in Fig. 31. According to the reasoning which shows the right-hand side of Eq. 496 to be a positive real constant, it follows from the similarity between the right-hand sides of Eqs. 496 and 505 that H is a positive real nonvanishing constant. This then is the condition for which the transformation 506 (or its equivalent 488) carries any circle through the fixed points z_1 and z_2 over into itself. This type of the transformation is designated as *hyperbolic*.

The second special type has the property that any one of the orthogonal family of circles enclosing the fixed points is transformed into itself. In order to recognize the condition which yields this result, one must recall that any circle enclosing the fixed points is, according to well-known principles in analytic geometry, defined as the locus of a point for which the ratio of its distances from the two fixed points (or poles) remains constant. If z denotes a variable point on such a circle,

$$\frac{|z - z_1|}{|z - z_2|} = \text{constant} \tag{507}$$

The condition that the transformed point ζ shall lie on the same circle evidently reads

$$\left|\frac{\zeta - z_1}{\zeta - z_2}\right| = \left|\frac{z - z_1}{z - z_2}\right| \tag{508}$$

which, with regard to Eq. 506, means that

$$|H| = 1 \qquad (H \neq 1) \tag{509}$$

The parenthetic statement $H \neq 1$ is intended again to rule out the trivial case $\zeta = z$. This special type of the transformation is designated as *elliptic*.

With regard to the further general properties of the linear fractional transformation given by Eqs. 488 or 506, it is useful to recognize that this transformation not only maps the points on a given circle in the z-plane upon a corresponding circle in the ζ-plane, but also furnishes a one-to one correspondence between all points within those regions of the z- and ζ-planes bounded by these circles. In other words, the transformation is said to map one of these regions upon the other.

Now a circle may be considered to be the boundary either of the region enclosed by it, or of the complementary region which lies outside it. In order to remove this ambiguity, and also to furnish a clearer visualization of how the mapping of corresponding regions is effected by the transformation, one may consider a specific circle K in the z-plane and the corresponding circle C in the ζ-plane. The points z_1, z_2, z_3 lie on the circle K, and the respectively numbered corresponding points ζ_1, ζ_2, ζ_3 lie on the circle C. In the order in which they are numbered, the points in each set form a sequence which fixes a definite direction of traversal (clockwise or counterclockwise) around their respective circle. In this way, corresponding reference directions of traversal are fixed for the two circles K and C. It is clear from the preceding discussion that a variable point following along the circle K in its reference direction is transformed into a point which follows along the circle C in the corresponding reference direction.

One may now imagine that at some point z the variable point following along K suddenly leaves this circular boundary by making a right-angle turn to the left. According to the principle of conformality, the corresponding variable point on C must likewise make a right-angle turn to the left at the point ζ corresponding to z. If one imagines a second circle K' in the z-plane lying wholly within the region to the left of K, with a diameter which differs from that of K by a small amount, this line of reasoning shows that the corresponding circle C' in the ζ-plane lies wholly within the region to the left of C. Continuing this process of reasoning (by applying the same line of thought to the circles K', C' and a pair K'', C'' lying wholly within the regions to the left of K' and C', etc.) one finds it clear that the entire region to the left (or to the right) of K is mapped upon the region to the left (or to the right) of the circle C.

If the reference directions of traversal around the circles K and C are the same (both clockwise or both counterclockwise), the region inside

(respectively outside) K is mapped upon the region inside (respectively outside) C, whereas if the reference directions are opposite, the region within K is mapped upon the region outside C, and vice versa.

As an illustrative example, let it be required to find the function which will map the region inside the unit circle about the origin upon the region outside the unit circle. Since the unit circle forms the common boundary for the two regions in question, one must look for that linear fractional transformation which transforms the unit circle into itself. If the reference direction of traversal for the unit circle in the z-plane is assumed to be counterclockwise, whereas for the unit circle in the ζ-plane it is assumed to be clockwise, the regions to the left of these boundaries in the z- and ζ-planes are those which are to be mapped upon each other. With these points in mind, the normal form for the required transformation is readily established.

In Eqs. 505 and 506 one may choose the following correspondence of points:

$$\begin{aligned} z_1 &= +1 \qquad & z_2 &= -1 \qquad & z_3 &= -j \\ \zeta_1 &= +1 \qquad & \zeta_2 &= -1 \qquad & \zeta_3 &= +j \end{aligned} \tag{510}$$

This makes $z = \pm 1$ correspond to the fixed points. Substituting into Eq. 505, one finds

$$H = \frac{j-1}{j+1} \bigg/ \frac{-j-1}{-j+1} = -1 \tag{511}$$

so that the normal form of the required transformation, according to Eq. 506, becomes

$$\frac{\zeta - 1}{\zeta + 1} = -\frac{z-1}{z+1} \tag{512}$$

This is equivalent to

$$\zeta = \frac{1}{z} \tag{513}$$

which is the simple transformation discussed earlier in this article, and evidently yields the desired mapping relationship. By choosing different sets of corresponding points which nevertheless satisfy the condition of opposite reference directions around the unit circle, one may find innumerable additional transformations which also map the inside of the unit circle upon the outside. It is, of course, not necessary that two of the chosen points be the fixed points of the transformation, since Eqs. 503 and 504 apply to any arbitrary sets of corresponding points.

As a second example, let it be required to find the substitution which

maps the inside of the unit circle in the z-plane upon the upper half of the ζ-plane. According to the principles discussed above, the following corresponding sets of points may be chosen:

$$\begin{array}{lll} z_1 = +1 & z_2 = -1 & z_3 = -j \\ \zeta_1 = +1 & \zeta_2 = -1 & \zeta_3 = 0 \end{array} \qquad [514]$$

The unit circle in the z-plane is traversed in the counterclockwise direction from $+1$ around to -1 and through $-j$ back to $+1$; the real axis in the ζ-plane is traversed from $+1$ through the point ∞ to -1 and thence through the origin back to $+1$. The regions to the left of these boundaries are those which are to be mapped upon each other. Substituting into Eq. 505, one finds

$$\mathrm{H} = \frac{-1}{1} \bigg/ \frac{-j-1}{-j+1} = \frac{1-j}{1+j} = -j \qquad [515]$$

so that the normal form of the desired transformation becomes, according to Eq. 506,

$$\frac{\zeta - 1}{\zeta + 1} = -j\frac{z-1}{z+1} \qquad [516]$$

This is equivalent to

$$\zeta = -j\frac{z+j}{z-j} \qquad \text{or} \qquad z = j\frac{\zeta - j}{\zeta + j} \qquad [517]$$

Here the points $z = \pm 1$ again are the fixed points.

Alternatively one may solve the present problem by assuming the corresponding sets of points

$$\begin{array}{lll} z_1 = +1 & z_2 = +j & z_3 = -1 \\ \zeta_1 = -1 & \zeta_2 = 0 & \zeta_3 = +1 \end{array} \qquad [518]$$

None of these are fixed points. Consequently the forms 503 and 504 must be used. According to Eq. 503,

$$\mathrm{H} = \frac{2}{1} \bigg/ \frac{-2}{-1-j} = 1 + j \qquad [519]$$

Substituting this into Eq. 504, one finds

$$\zeta = -j\frac{z-j}{z+j} \qquad \text{or} \qquad z = -j\frac{\zeta - j}{\zeta + j} \qquad [520]$$

Comparison with Eq. 517 shows that these results are identical except that z is replaced by $-z$.

From the nature of this problem it is obvious that Eqs. 517 or 520 still

constitute solutions if z is replaced by $ze^{j\phi}$, where ϕ is any chosen angle. For example, replacing z in Eq. 517 by $-jz$ yields another possible solution, namely,

$$\zeta = j\frac{1-z}{1+z} \qquad \text{or} \qquad z = \frac{j-\zeta}{j+\zeta} \tag{521}$$

This has the corresponding points

$$\begin{aligned} z &= +1, \quad +j, \quad -1, \quad -j \\ \zeta &= \ \ 0, \quad +1, \quad \infty, \quad -1 \end{aligned} \tag{522}$$

Again none of these are fixed points. The point $z = -1$ here corresponds to the point ∞ in the ζ-plane. In the transformation 517 the latter corresponds to the point $z = j$; in Eq. 520 the point $z = -j$ corresponds to $\zeta = \infty$. Evidently the present problem may be solved with any desired point on the unit circle in the z-plane corresponding to the point ∞ in the ζ-plane.

Alternatively, one may for example map the right half of the ζ-plane upon the inside of the unit circle in the z-plane. The transformation which yields this result is obtained from any of the solutions to the preceding example by replacing ζ by $j\zeta$. This change amounts to making the substitution $\zeta' = -j\zeta$ and subsequently writing ζ for ζ' again (a substitution which rotates all points by 90 degrees in the negative direction). Applied to the Eqs. 521, for example, this process yields

$$\zeta = \frac{1-z}{1+z} \qquad \text{or} \qquad z = \frac{1-\zeta}{1+\zeta} \tag{523}$$

for which a set of corresponding points are

$$\begin{aligned} z &= +1, \quad +j, \quad -1, \quad -j \\ \zeta &= \ \ 0, \quad -j, \quad \infty, \quad +j \end{aligned} \tag{524}$$

It is observed that these are the same points as those in the set 522 except that the ζ-points are rotated 90 degrees in the negative direction, so that the imaginary instead of the real axis becomes the boundary of the mapped region in the ζ-plane.

It is instructive to study the conformal maps of one of these functions in somewhat greater detail. The transformation 523 is particularly interesting in this respect, since it is perfectly symmetrical in the variables z and ζ. Figure 33 shows how a system of concentric circles about the origin in the z-plane, along with the orthogonal family of radial lines, are mapped in the ζ-plane. The origin in the z-plane becomes the point $+1$ in the ζ-plane. The concentric circles within the unit circle of the z-plane become eccentric circles about $+1$ in the ζ-plane, the unit circle

itself corresponding to the imaginary axis of the ζ-plane. The radial lines within the unit circle of the z-plane become those portions of the orthogonal circles through the points ± 1 which lie in the right half of the ζ-plane. The circles of the z-plane (shown dotted) which are larger than the unit circle become eccentric circles in the ζ-plane about the point -1, and the portions of the orthogonal circles through the points $\zeta = \pm 1$ which lie in the left half of that plane (also shown dotted) correspond to those portions of the radial lines in the z-plane lying outside the unit circle.

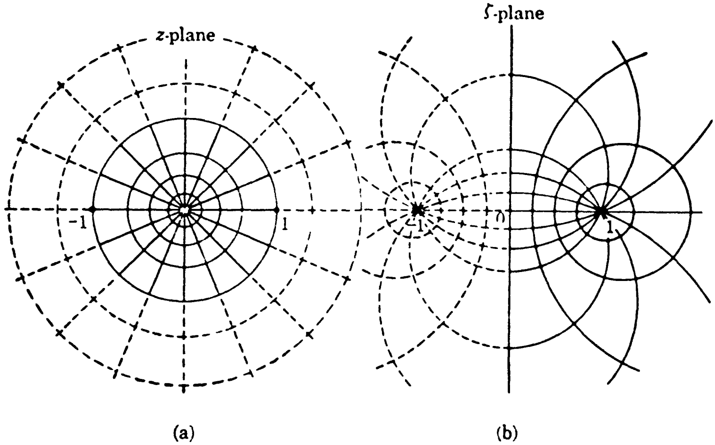


FIG. 33. Map of the interior of the unit circle in the z-plane to the right-half ζ-plane by the transformation $z = (1 - \zeta)/(1 + \zeta)$.

A more comprehensive view of this situation is gained through recognition that the points $+1$ and -1 of the ζ plane are transformed into the origin (south pole) and the point ∞ (north pole), respectively, of the complex sphere associated with the z-plane. In other words, the poles of the complex sphere associated with the z-plane are transformed into the finite points ± 1 of the ζ-plane, and the resulting map in that plane represents the corresponding distortion suffered by the concentric circles and radial lines in the z-plane which accompanies this shifting of the poles. This interpretation is responsible for the term " bipolar circles " or " bipolar plot " by which the map in the ζ-plane is also known.

Since the forms 523 are symmetrical in the variables z and ζ, it follows that the sets of orthogonal loci in the z- and ζ-planes of Fig. 33 may be interchanged. That is, if concentric circles about the origin together with the family of radial lines are drawn in the ζ-plane, the corresponding map in the z-plane becomes that which in Fig. 33 is shown for the ζ-plane. In other words, the transformation 523 may alternatively be said to map

the inside of the unit circle in the ζ-plane upon the right half of the z-plane.

A word of caution may be appropriate at this point in order to guard the reader against confusing part (b) of Fig. 33 with the similar appearing plot of Fig. 32. In the latter, the poles z_1 and z_2 are the fixed points of the transformation. Figure 33(b), on the other hand, is *not* the corresponding plot for the transformation 523, since the fixed points in this case are evidently not the points $\zeta = \pm 1$. The maps in Fig. 33 are merely a pair of corresponding sets of curves chosen to illustrate the way in which the region enclosed by the unit circle is mapped upon the right half plane, and these curves have nothing in common with the type of plot shown in Fig. 32. The reader may further clarify his thoughts on this score by determining the fixed points of the transformation 523 and by subsequently drawing loci of the type shown in Fig. 32.

In connection with the preceding general discussion it is relevant to add the following remarks regarding several collaterally useful properties of the linear fractional transformation. It should readily be appreciated, for example, that the end result of carrying out two different linear fractional transformations in succession can also be had from a single one. More specifically, if the relation

$$z' = \frac{a_1 z + b_1}{c_1 z + d_1} \tag{525}$$

represents a transformation from the variable z to a new variable z', and

$$z'' = \frac{a_2 z' + b_2}{c_2 z' + d_2} \tag{526}$$

represents a succeeding transformation to still another variable z'', it is always possible to relate z'' to z directly by a transformation of the same form, namely,

$$z'' = \frac{az + b}{cz + d} \tag{527}$$

The truth of this statement follows immediately from the consideration that the linear fractional form carries circles over into circles and conversely that any univalued analytic transformation of circles into circles is linear fractional (see footnote on page 375). Thus, two such transformations applied in tandem obviously accomplish no more than can be accomplished by a single one.*

*Because of this property the totality of possible linear fractional transformations is said to form a *group*.

The coefficients a, b, c, d in the single resulting transformation 527 may be determined from the coefficients in the two separate transformations 525 and 526, through substituting the relation 525 for z' into Eq. 526 and putting the result into the form of Eq. 527. One finds that

$$\begin{aligned} a &= a_2a_1 + b_2c_1 \\ b &= a_2b_1 + b_2d_1 \\ c &= c_2a_1 + d_2c_1 \\ d &= c_2b_1 + d_2d_1 \end{aligned} \qquad [528]$$

It is interesting (as well as relevant in dealing with circuit problems involving a cascade of transmission networks) to observe that the relations 528 may be expressed more compactly by means of the single matrix equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \times \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \qquad [529]$$

Another interesting fact is brought out when the linear fractional transformation, Eq. 488, is put into the form

$$\zeta = \frac{a}{c} + \frac{\left(\frac{b}{c} - \frac{ad}{c^2}\right)}{z + \frac{d}{c}} \qquad [530]$$

as is always possible so long as the conditions 489 and 490 are fulfilled. In view of this form, one may regard the arbitrary linear fractional transformation as equivalent to carrying out in succession the simpler transformations indicated by

$$z' = z + \frac{d}{c} \qquad [531]$$

$$z'' = \frac{1}{z'} \qquad [532]$$

$$\zeta = \left(\frac{b}{c} - \frac{ad}{c^2}\right) z'' + \frac{a}{c} \qquad [533]$$

The first of these component transformations represents a displacement, the second is a simple inversion, and the third represents a magnification and rotation (multiplication by a complex constant) followed by a displacement. Since each of these component transformations carries circles over into circles, one arrives again at the conclusion that the linear fractional transformation has this property.

Finally, an interesting correlation with the process of stereographic projection may be mentioned. As pointed out in Art. 4, this geometrical process is one whereby every finite point in the complex z-plane has uniquely associated with it a point on the so-called complex sphere which is tangent at the origin. If a second plane (ζ-plane) is imagined to be tangent to the same sphere at some other point, and if the same geometrical process is used to associate points on the sphere with points in the second plane, it is evidently possible to state that the system consisting of the two planes and the sphere, together with the geometrical process of stereographic projection, enables one to associate uniquely a point z in one of these planes with a corresponding point ζ in the other plane.

It is of interest to show how the transformation implied by this process can be expressed analytically in terms of the linear fractional function 488.* Although the converse of this statement is not generally true, it is worth noting that some of the linear fractional transformations encountered in practical problems may be interpreted geometrically in this simple fashion.

In order to develop the appropriate analytic relationships, one may begin by visualizing a sphere of diameter D and any two tangent planes (the z- and ζ-planes). The two points of tangency determine uniquely a great circle on the sphere. The plane containing this great circle is conveniently chosen as a cross-sectional plane in which to indicate graphically further pertinent geometrical relationships. Figure 34 shows the system as viewed in this plane, to which the z- and ζ-planes are both orthogonal and hence appear as lines. The two polar axes, which are diameters of the sphere and are normal to the z- and ζ-planes at their respective points of tangency, make an angle with each other which is denoted by γ. The real and imaginary axes in the z- and ζ-planes have orientations which at the moment may be considered to be arbitrary.

The points in the ζ-plane corresponding to $z = 0$ and $z = \infty$ are denoted by ζ_0 and ζ_∞, respectively, whereas those in the z-plane corresponding to $\zeta = 0$ and $\zeta = \infty$ are denoted by z_0 and z_∞. These four points lie in the plane of the polar axes, and their locations in the z- and ζ-planes are shown in Fig. 34. From the geometry it is readily seen that

$$|z_0| = |\zeta_0| = D \tan \frac{\gamma}{2} \qquad [534]$$

*The possibility of doing so should be evident from the geometrical proposition that the process of stereographic projection carries circles in the plane over into circles on the sphere, and vice versa. Hence a circle in one of the planes is transformed into a circle in the other plane. Any univalued analytic transformation which carries circles over into circles is expressible in terms of a linear fractional function.

and

$$|z_\infty| = |\zeta_\infty| = D \cot \frac{\gamma}{2} \qquad [535]$$

in which the absolute value signs are necessary because the lines representing the z- and ζ-planes in Fig. 34 are not necessarily coincident with the real axes in these planes.

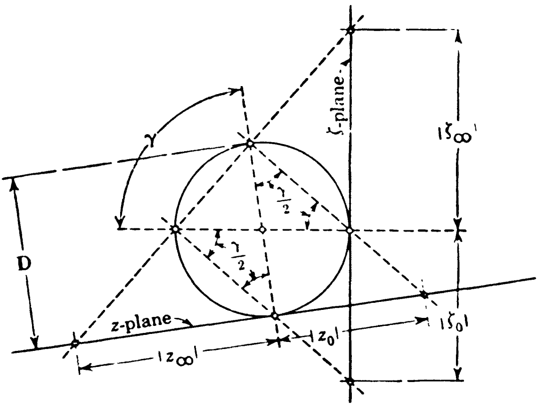


FIG. 34. In many cases, a linear fractional transformation may be visualized as a twofold stereographic projection in terms of two planes tangent to the same complex sphere.

With reference to the linear fractional function 488 and its inverse as given by Eq. 492, it is expedient in the subsequent considerations to observe that only three of the four constants a, b, c, d are independent, since the right-hand members of these relations have the same value if their numerators and denominators are multiplied by a common (finite, nonzero) factor. Inasmuch as it is assumed that $c \neq 0$, one may choose the constant $1/c$ as this common factor, or, what amounts to the same thing, arbitrarily set c equal to unity. Equations 488 and 492 then yield the results

$$z_0 = -\frac{b}{a} \qquad z_\infty = -d \qquad [536]$$

and

$$\zeta_0 = \frac{b}{d} \qquad \zeta_\infty = a \qquad [537]$$

Comparison with Eqs. 534 and 535 shows that

$$|a| = |d| = D \cot \frac{\gamma}{2} \tag{538}$$

and

$$|b| = D^2 \tag{539}$$

Referring to Fig. 34, one recognizes that the angles of the complex quantities z_0 and z_∞ (in the complex z-plane) must differ by π radians and that the same is true for the quantities ζ_0 and ζ_∞. In view of Eqs. 536 and 537, this fact yields the conclusion that

$$\frac{b}{ad} \quad \text{must be a negative real number} \tag{540}$$

This result, together with the condition that $|a| = |d|$, represent the restrictions which must be imposed upon a given linear fractional function in order that it may possess the simple geometrical interpretation discussed here. For any given linear fractional function fulfilling these conditions, the diameter of the sphere and the angle γ between the polar axes are given by Eqs. 538 and 539, whereas the orientations of the real and imaginary axes in the z- and ζ-planes are obtained from Eqs. 536 and 537. Thus the corresponding geometrical configuration is determined.

Conversely, if the geometrical configuration for the sphere and the two tangent planes is given, and one wishes to find the corresponding linear fractional function, three of the relations contained in Eqs. 536 and 537 may be used to determine the complex constants a, b, and d. The magnitudes of these constants must, of course, agree with the values given by Eqs. 538 and 539. This determination is always possible.

As an illustration it is interesting to consider the transformation 523 by means of which the interior of the unit circle in either plane (z or ζ) is mapped upon the right half of the other plane. By inspection of Eqs. 523 one sees that

$$a = -1 \qquad b = 1 \qquad d = 1 \tag{541}$$

and also that

$$\begin{aligned} z_0 &= 1 \qquad z_\infty = -1 \\ \zeta_0 &= 1 \qquad \zeta_\infty = -1 \end{aligned} \tag{542}$$

These values are consistent with Eqs. 536 and 537 as, of course, they should be. From Eqs. 538 and 539 one obtains

$$D = 1 \qquad \gamma = \frac{\pi}{2} \tag{543}$$

The transformation in question is, therefore, given geometrically (according to the present interpretation) by means of a sphere of unit diameter and a pair of tangent planes which are normal to each other as well as normal to the plane containing the polar axes. This configuration is shown in Fig. 35. Although the values 542 uniquely determine the relative orientations and positive directions for the real axes in the z- and ζ-planes, the positive directions for the imaginary axes are not uniquely fixed, because all these values are real. If the imaginary axis for the z-plane is assumed to point vertically upward from the plane of the paper in Fig. 35, and the unit circle in the z-plane is traversed in the counterclockwise direction so as to keep the interior on the left, then, since the imaginary axis in the ζ-plane must be traversed in such a direction as to keep the right half plane on the left, one observes that the imaginary axis for the ζ-plane points vertically into the plane of the paper in Fig. 35. This choice for the positive directions of the imaginary axes places the top faces of the z- and ζ-planes opposite the sphere. It is alternatively possible to assume both positive directions for the imaginary axes reversed, whence the top faces of the planes appear *adjacent* to the sphere.

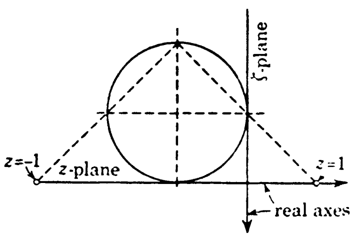


FIG. 35. Visualization of the transformation $z = (1 - \zeta)(1 + \zeta)$ as a twofold stereographic projection.

To anyone who has developed a facility for visualizing geometrical objects in three dimensions, this interpretation for the transformation 523 affords a useful means for correlating its various detailed characteristics.

25. A more general mapping function; the Schwarz-Christoffel formula

The principal field of usefulness with regard to conformal mapping is found in connection with problems in potential theory. When the geometry of the physical system exhibits longitudinal uniformity in one of its dimensions, the problem reduces to a two-dimensional one. As shown in Art. 5, the real and imaginary parts u and v of a function of a complex variable satisfy Laplace's equation in two dimensions and hence represent potential functions. Since the loci for u = constant and v = constant form orthogonal families, they may be regarded as the equipotential and flow lines of a nonturbulent field, that is, of a pure potential field. If a pair of such functions can be found whose loci conform to the geometrical boundaries of a given physical system, the solution to the boundary value problem for that system is thereby given.

For example, Poisson's integral, Eq. 323 or Eq. 324, yields the potential inside a circle in terms of specified boundary values. This integral, therefore, constitutes a solution to the boundary value problem for the circle, the specific forms given by Eqs. 332 and 333 expressing this solution as a sum of exponential functions.

By means of a function which conformally maps the interior of a circle upon a region having a different geometrical configuration, the boundary value problem for that configuration may be transformed into an equivalent one for the circle, and the desired solution is found through applying to the solution for the circle, the inverse transformation.

The boundary is commonly an equipotential locus, so that the specification of boundary values amounts merely to stipulating that the potential function shall be constant on the boundary. In electrical problems, this boundary is the surface of a conductor or of a material having a large permeability relative to that of the surrounding medium. The solution to the boundary value problem for the circle then consists simply of concentric circles for the equipotential loci, with the orthogonal radial lines representing the flow field. The origin, or center of the circle, and the point at infinity represent the source and sink for this field.

An even simpler geometrical configuration is given by a pair of boundaries in the form of parallel lines. When these are constant potential loci, the flow map consists of a rectangular grid of flow lines and equipotential lines. Such a rectangular region is mapped upon the interior of the circle by means of the function

$$z = e^t \qquad \text{or} \qquad t = \ln z \tag{544}$$

Writing

$$z = re^{j\phi} \qquad \text{and} \qquad t = m + jn \tag{545}$$

one finds from Eq. 544 that

$$m = \ln r \qquad \text{and} \qquad n = \varphi \tag{546}$$

If the origin ($r = 0$) and infinity ($r = \infty$) in the z-plane represent the source and sink for the field, the radial flow enclosed by a pair of lines $\phi = \phi_1$ and $\phi = \phi_2$ (these may be $\phi = 0$ and $\phi = 2\pi$) is mapped upon the t-plane as a parallel flow enclosed by the straight lines $n = \phi_1$ and $n = \phi_2$. The equipotential loci, which in the z-plane are the circles $r =$ constant, become the straight lines $m =$ constant, which are at right angles to the flow lines $n =$ constant. Source and sink become the points $m = \pm\infty$. These matters are illustrated in Fig. 36. The rectangular flow map in the t-plane may be regarded as a reference field (a flow map of the simplest type) to which the radial type of flow map is transformed by the function 544.

Inasmuch as the upper half of the z-plane may be conformally mapped upon the upper half of the ζ-plane, according to the discussion of the previous article, with any two desired points* on the real axis of one half plane corresponding to the origin and infinity in the other, it follows that a flow map in the upper half plane, with any two points on the real axis designated as the source and sink, may be reduced to the reference field in the t-plane of Fig. 36. If transformations can be found which map regions having other geometrical configurations for their boundaries, upon the upper half plane, or upon the region enclosed by the unit circle, a way is established for also reducing the flow maps for these configurations to a simple rectangular reference field, thus making possible the solution to boundary value problems in these more complicated cases.

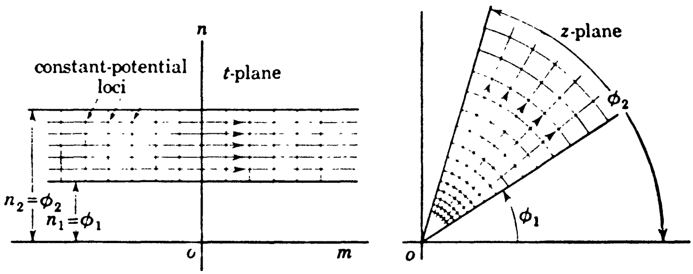


FIG. 36. Transformation of a radial flow map to a simpler flow pattern by the transformation $z = e^t$.

An extremely useful mapping function, of considerable generality in its ability to meet various geometrical configurations, is given by the so-called *Schwarz-Christoffel formula*, which reads

$$w(z) = M\int_{z_0}^{z} (\zeta - z_1)^{-\mu_1}(\zeta - z_2)^{-\mu_2}\cdots(\zeta - z_n)^{-\mu_n}\,d\zeta + N \qquad [547]$$

Here ζ is a running variable in the z-plane, $z_1, z_2, \cdots z_n$ are n finite points on the real axis, numbered in such an order that

$$z_1 < z_2 < \cdots < z_n \qquad [548]$$

and the quantities $\mu_1, \mu_2, \cdots \mu_n$ appearing in the exponents are any set of positive or negative real numbers. The constants M and N may have complex values, with the possibility that N be zero, but M must, of course, have a nonzero value. The lower limit z_0 of the integral is an arbitrary point in the upper half plane. It may be chosen equal to zero, or equal to one of the points $z_1 \cdots z_n$. The principle guiding this choice is best seen from the illustrative examples given subsequently.

*A third point (for example, midway between the other two) may be chosen to correspond to the point +1 (which lies between the origin and infinity).

The independent variable for the mapping function $w(z)$ is the upper limit of the integral. For this reason the derivative of the function is given by

$$\frac{dw}{dz} = M(z - z_1)^{-\mu_1}(z - z_2)^{-\mu_2} \cdots (z - z_n)^{-\mu_n} \qquad [549]$$

as may be seen from the fact that if one has

$$w(z) = \int_{z_0}^{z} f(\zeta)\, d\zeta \qquad [550]$$

the usual definition for the derivative

$$\frac{dw}{dz} = \underset{\Delta z \to 0}{\text{limit}} \left[\frac{w(z + \Delta z) - w(z)}{\Delta z}\right] \qquad [551]$$

yields

$$w(z + \Delta z) - w(z) = \int_{z}^{z+\Delta z} f(\zeta)\, d\zeta \qquad [552]$$

Since Δz is a small displacement (becoming zero in the limit), one may say that for the integration in Eq. 552 the function $f(\zeta)$ is essentially constant and equal to the value $f(z)$. It is assumed, of course, that the function $f(\zeta)$ is continuous in the vicinity of the point $\zeta = z$, which is a recognized condition for the existence of the derivative in the first place. With $f(\zeta)$ equal to the constant value $f(z)$, it may be placed in front of the integral sign, and Eq. 552 yields

$$w(z + \Delta z) - w(z) \cong f(z) \int_{z}^{z+\Delta z} d\zeta = f(z)\, \Delta z \qquad [553]$$

the approximation becoming exact in the limit $\Delta z \to 0$. Completing the limit, one finds, therefore, that

$$\frac{dw}{dz} = f(z) \qquad [554]$$

The essential character of the function $w(z)$ may now be recognized from a study of the behavior of the derivative 549 in the vicinity of the points $z = z_\nu$. The first step in this direction is to represent the factor $(z - z_\nu)$ in the polar form as illustrated in Fig. 37. This representation reads

$$(z - z_\nu) = |z - z_\nu| e^{j(\phi_\nu + 2\pi k)} \qquad [555]$$

in which k is an integer.
Then

$$(z - z_\nu)^{-\mu_\nu} = |z - z_\nu|^{-\mu_\nu} e^{-j(\mu_\nu\phi_\nu + 2\pi k\mu_\nu)} \qquad [556]$$

Since the quantity μ_ν is not necessarily an integer, the right-hand side of Eq. 556 may have many different values for different integer values of k. In order to remove this multivaluedness of the factor $(z - z_\nu)^{-\mu_\nu}$, it is specified at the outset that k shall assume only the value zero. This specification is equivalent to stating that the function dw/dz is to be studied on only one of the many leaves of its Riemann surface, namely, on that one which corresponds to $k = 0$ in Eq. 556. A typical factor in Eq. 549 then becomes

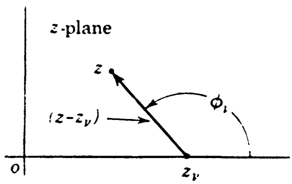


FIG. 37. Representation of $z - z_\nu$ in polar form in the study of dw/dz.

$$(z - z_\nu)^{-\mu_\nu} = |z - z_\nu|^{-\mu_\nu} e^{-j\mu_\nu\phi_\nu} \quad [557]$$

and if the point z is allowed to lie only in the upper half plane or on the real axis of the z-plane, it is clear from Fig. 37 that

$$0 \leqq \varphi_\nu \leqq \pi \quad [558]$$

When the polar forms

$$M = |M| e^{j\alpha} \quad [559]$$

and

$$\frac{dw}{dz} = \left|\frac{dw}{dz}\right| e^{j\theta} \quad [560]$$

are introduced, it follows that

$$\theta = \alpha - \mu_1\phi_1 - \mu_2\phi_2 - \cdots - \mu_n\phi_n \quad [561]$$

It is now assumed that the variable z in the function 549 is restricted to real values only; that is, the variable point z is thought of as moving along

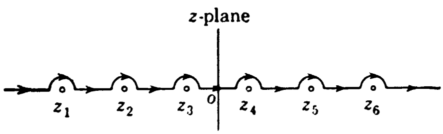


FIG. 38. The path along which dw/dz is studied in the Schwarz-Christoffel transformation.

the real axis from $-\infty$ to ∞, the only deviation from this behavior occurring wherever the variable point z encounters one of the critical points z_ν. There it makes a slight detour around the critical point instead of passing directly through it. These detours may be visualized as having the form of vanishingly small semicircular arcs lying in the upper half plane, as shown in Fig. 38. As the point z traverses a small semicircular arc in the vicinity of the point z_ν, the angle ϕ_ν changes from the value π to zero,

whereas the angles of the remaining factors do not change at all because of the assumed vanishingly small radius of the semicircular detour. Hence for the range

$$z_{\nu-1} < z < z_{\nu+1} \quad [562]$$

one has

$$\phi_1 = \phi_2 = \cdots = \phi_{\nu-1} = 0 \qquad \pi \geqq \phi_\nu \geqq 0 \qquad \phi_{\nu+1} = \phi_{\nu+2} = \cdots = \phi_n = \pi \quad [563]$$

and* according to Eq. 561

$$\alpha - (\mu_\nu + \mu_{\nu+1} + \cdots + \mu_n)\pi \leqq \theta \leqq \alpha - (\mu_{\nu+1} + \mu_{\nu+2} + \cdots + \mu_n)\pi \quad [564]$$

Throughout the range 562, the angle θ is, therefore, increased by the amount

$$\Delta\theta = \mu_\nu \pi \quad [565]$$

the important feature being that this increment occurs only as the point z traverses the small semicircular arc. In other words, as the point z moves along the real axis, the angle θ remains constant as z proceeds from one of the critical points to the next, receiving a sudden increment $\Delta\theta = \mu_\nu\pi$ only as z passes directly over the critical point z_ν.

According to the discussion of conformal mapping in Art. 2, it is recognized that the map of the function $w(z)$ in the w-plane, corresponding to the real axis in the z-plane, consists of a succession of straight-line segments between the points w_1, w_2, $\cdots$ corresponding respectively to $z_1, z_2, \cdots$, the angular directions of two consecutive segments confluent in the point w_ν differing by $\mu_\nu\pi$. That is, the map in the w-plane of the function 547, corresponding to the real axis in the z-plane, traversed from $-\infty$ to ∞, has the general character shown in Fig. 39. This result follows from the fact, pointed out in Art. 2, that the angle of dw/dz equals the difference between the angles of the increments dw and dz, and since the angle of the latter remains zero as the point z moves along the real axis, the angle of dw/dz must equal that of dw. This angle, however, is shown to remain constant except when z passes over one of the critical quantities z_ν. At the corresponding points w_ν then, the direction of the increment dw suddenly changes by the amount $\mu_\nu\pi$.

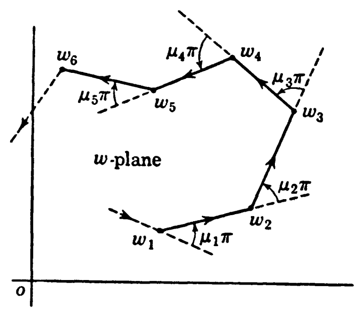


FIG. 39. The map in the w-plane of the real axis in the z-plane shown in Fig. 38.

*If μ_ν is negative, the inequalities in Eq. 564 are reversed.

The plot in the w-plane corresponding to the real axis in the z-plane is thus seen to be a polygon with the points $w_1 \cdots w_n$ as its vertexes. If

$$\mu_1 + \mu_2 + \cdots + \mu_n = 2 \qquad [566]$$

the sum of the increments $\Delta\theta$ at the n vertexes $w_1 \cdots w_n$ equals 2π. Since 2π equals the sum of the external angles of a closed polygon, it follows that when the condition 566 is fulfilled, the polygon is n-sided. Unless one or more of the values of $\mu_1 \cdots \mu_n$ are equal to or greater than unity, all the n vertexes of the polygon lie at finite points corresponding to the finite values $z_1 \cdots z_n$ on the real axis of the z-plane. The point at infinity in the z-plane corresponds to an ordinary point on the straight-line segment joining the vertexes w_n and w_1. If a factor μ_ν has a positive value equal to or greater than unity, the corresponding vertex of the polygon in the w-plane lies at infinity. This circumstance is subsequently discussed in detail.

When the condition 566 is not fulfilled, the point at infinity in the z-plane also corresponds to a vertex, and the polygon in the w-plane has $n + 1$ sides. In order to consider this possibility more fully it is necessary first to study the function $w(z)$ in the vicinity of the point $z = \infty$. For this purpose it is convenient to make the substitution

$$\zeta^* = \frac{-1}{\zeta}, \quad \text{with} \quad z^*_\nu = \frac{-1}{z_\nu} \qquad [567]$$

Then

$$(\zeta - z_\nu)^{-\mu_\nu} = (\zeta^* z^*_\nu)^{\mu_\nu} \cdot (\zeta^* - z^*_\nu)^{-\mu_\nu} \qquad [568]$$

and

$$d\zeta = \frac{d\zeta^*}{(\zeta^*)^2} \qquad [569]$$

so that Eq. 547 takes the form

$$w(z^*) = M \int_{z^*_0}^{z^*} \frac{(z^*_1)^{\mu_1} \cdots (z^*_n)^{\mu_n} (\zeta^*)^{(\mu_1 + \mu_2 + \cdots + \mu_n - 2)}}{(\zeta^* - z^*_1)^{\mu_1} (\zeta^* - z^*_2)^{\mu_2} \cdots (\zeta^* - z^*_n)^{\mu_n}} \, d\zeta^* + N \qquad [570]$$

in which

$$z^* = \frac{-1}{z} \qquad [571]$$

The derivative reads

$$\frac{dw}{dz^*} = \frac{M (z^*_1)^{\mu_1} \cdots (z^*_n)^{\mu_n} (z^*)^{(\mu_1 + \mu_2 + \cdots + \mu_n - 2)}}{(z^* - z^*_1)^{\mu_1} (z^* - z^*_2)^{\mu_2} \cdots (z^* - z^*_n)^{\mu_n}} \qquad [572]$$

The point $z = \infty$ corresponds to $z^* = 0$. For the vicinity of $z^* = 0$,

Eq. 572 gives

$$\frac{dw}{dz^*} \cong (-1)^n M(z^*)^{(\mu_1+\mu_2+\cdots+\mu_n-2)} \qquad [573]$$

Hence it is clear that the point $z = \infty$ yields a vertex at which the angular increment $\Delta\theta$ has the value

$$\Delta\theta = (2 - \mu_1 - \mu_2 - \cdots - \mu_n)\pi \qquad [574]$$

This vertex also lies at a finite point in the w-plane unless the quantity $(2 - \mu_1 - \mu_2 - \cdots - \mu_n)$ is equal to or larger than unity. It is obvious that the vertex does not exist if the condition 566 is fulfilled.

In order to study the behavior of the function w in the vicinity of one of its vertexes, one may first write for the derivative 549

$$\frac{dw}{dz} = (z - z_\nu)^{-\mu_\nu}[a_0 + a_1(z - z_\nu) + a_2(z - z_\nu)^2 + \cdots] \qquad [575]$$

in which the bracket expression is a Taylor expansion for the function 549 with the factor $(z - z_\nu)^{-\mu_\nu}$ missing. For this Taylor expansion, which is evidently possible because the function it represents is regular at the point $z = z_\nu$, the coefficient a_0 is certainly not zero. Hence assuming $\mu_\nu < 1$ the term by term integration of Eq. 575 yields

$$w = \int_{z_0}^{z} dw = \frac{a_0(z - z_\nu)^{-\mu_\nu+1}}{-\mu_\nu + 1} + \frac{a_1(z - z_\nu)^{-\mu_\nu+2}}{-\mu_\nu + 2} + \cdots + w_\nu \qquad [576]$$

For the immediate vicinity of the vertex at the point $w = w_\nu$, therefore, one has the representation

$$w \cong w_\nu + \frac{a_0(z - z_\nu)^{-\mu_\nu+1}}{1 - \mu_\nu} \qquad [577]$$

This analysis shows that so long as $\mu_\nu < 1$, the function $w(z)$ is regular in the vicinity of the vertex w_ν. This vertex may have a variety of ap-

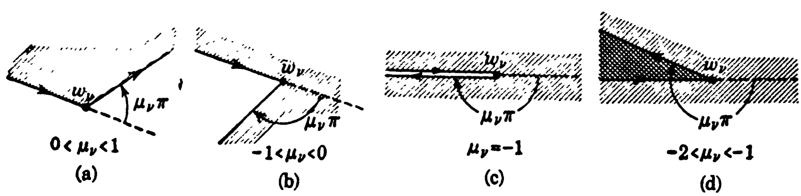


FIG. 40. Appearance of vertexes of polygon in the w-plane with varying values of μ.

pearances depending upon the particular value of $\mu_\nu < 1$. Some of these are shown in Fig. 40. The arrows on the line segments indicate the direc-

tion in which the contour of the polygon is traversed as the point z travels in the positive direction (from $-\infty$ to $+\infty$) along the real axis of the z-plane. The region enclosed by the polygon, which is that on the left of the contour, is shown shaded.

Part (a) of the figure shows the appearance of the vertex when $0 < \mu_\nu < 1$. Here the external angle has a positive value between zero and π. Part (b), for which the external angle has a negative value between zero and $-\pi$, corresponds to $-1 < \mu_\nu < 0$. Part (c) shows the vertex for $\mu_\nu = -1$. Here the external angle equals $-\pi$. Finally, part (d) illustrates the appearance of a vertex at which the external angle is less than $-\pi$ (that is, equal to the negative of a quantity which is larger than π). In this case the enclosed region involves a double-mapped portion.

If $\mu_\nu = 1$, Eq. 575 yields

$$\frac{dw}{dz} = a_0(z - z_\nu)^{-1} + a_1 + a_2(z - z_\nu) + \cdots \qquad [578]$$

and the integration then gives

$$w = \int_{z_0}^{z} dw = a_0 \ln (z - z_\nu) + a_1(z - z_\nu) + \frac{a_2}{2}(z - z_\nu)^2 + \cdots + C \quad [579]$$

in which

$$C = -a_0 \ln (z_0 - z_\nu) - a_1(z_0 - z_\nu) - \cdots \qquad [580]$$

In this case the vertex w_ν evidently lies at infinity, since the value of the right-hand side in Eq. 579 becomes infinite for $z = z_\nu$. Figure 41 shows how this vertex may be imagined to appear, if it is assumed that one is permitted to indicate infinity as a finite point. The enclosed region in this vicinity is null, since it is contained between line segments which fall upon each other. In the finite region of the w-plane the appearance of a polygon with a vertex w_ν of this sort at infinity is shown in Fig. 42. The vertex of Fig. 41 at infinity is seen to result from a pair of confluent line segments which are parallel and hence meet or intersect at infinity.

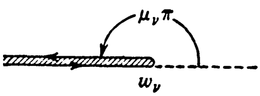


FIG. 41. Appearance of the vertex when $\mu_\nu = 1$.

When μ_ν has any value larger than unity, the corresponding vertex also lies at infinity. Figure 43 shows the appearance of a polygon with such a vertex. The external angle lies between π and 2π, and μ_ν has a value between 1 and 2. For $\mu_\nu = 2$ the external angle equals 2π. Analysis similar to the preceding then shows that the function $w(z)$ has a simple pole at the point $z = z_\nu$ in addition to having a logarithmic infinity there.

In the light of the preceding discussion and interpretation it may now

be stated by way of a summary that the function defined by the integral 547 uniquely and continuously maps all points on the real axis of the z-plane upon the boundary of a polygon in the w-plane. Since its derivative as expressed by Eqs. 549 and 572 is regular at all points in the upper half of the z-plane inclusive of the point at infinity, it follows that this function is regular and continuous throughout this entire half plane and is single-valued by virtue of the stipulation regarding the choice of values for the factors $(\zeta - z_\nu)^{-\mu_\nu}$. It may be concluded, therefore, that the function also maps the entire upper half of the z-plane upon the region en-

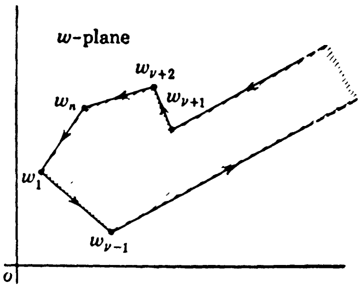


FIG. 42. Appearance of the polygon in the finite w-plane when one of its vertexes corresponds to $\mu_\nu = 1$.

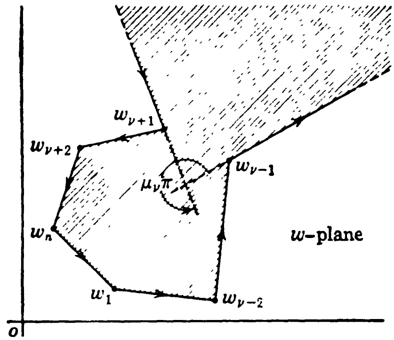


FIG. 43. Appearance of the polygon in the finite w-plane when one of its vertexes corresponds to $\mu_\nu > 1$.

closed by the polygon in the w-plane, for any closed boundary lying wholly inside the polygon must by reason of the continuity and single valuedness of the function also lie wholly within the upper half of the z-plane. The availability of this mapping function and that represented by the linear fractional form discussed in the preceding article greatly enhances the means for solving two-dimensional boundary value problems, as is illustrated by the following examples.

The first problem to be discussed is the determination of the field distribution in the vicinity of the edges of a parallel-plate condenser. Since the distortion of the field ("fringing") is confined to the more immediate vicinity of the edge, the plates may be assumed to be infinitely wide, and since the field distribution is symmetrical about a center line between the plates, it is sufficient to map the field on one side of this center line only. The region over which a field map is desired may, therefore, be sketched as shown in Fig. 44. The edge of the condenser plate is at the point C above the origin of the w-plane. The plate itself is parallel to the real axis of this plane and extends infinitely far to the left. The real axis

represents the center line between the plates, so that the distance d equals half the spacing between them.

The shaded area, throughout which the field map extends, is regarded as the region enclosed by a polygon which has three vertexes. One of these lies at the finite point C, and the other two lie at infinity. One of these latter two results from the region A with its parallel boundaries extending to infinity on the left. This one, which may be designated as "vertex A," evidently has the character of the vertex at infinity for the polygon of Fig. 42. The other vertex at infinity, which may be designated as "vertex B," has the character of the vertex w_ν in Fig. 43 for $\mu_\nu = 2$. It represents the infinite extension of the region B to the right and upper left of Fig. 44.

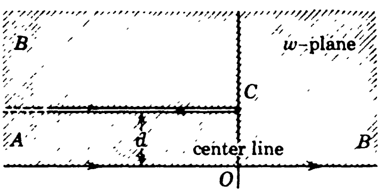


Fig. 44. Relevant to determining the field of a parallel-plate condenser.

The electric flux being considered as the fluid, the flow in the parallel-plate condenser is from one of its plates to the other, whereas the equipotential loci are the orthogonal curves symmetrically grouped about the center line. However, from a mathematical standpoint it is equally admissible to consider the flow in the space between the plates to be in the general direction along the center line, with the equipotential loci extending from one plate to the other. In other words, the flow map consisting of orthogonal families of loci depends only upon the geometrical configuration of the boundaries, and hence it makes no difference which of the families of curves is thought of as representing the flow of a physical fluid. In the present instance it is convenient to think of the fluid as flowing sideways, with the plates as longitudinal boundaries and the vicinity of the edges as the throat from which this fluid issues into the entire surrounding space. According to this point of view, the vertex A becomes the source and the vertex B becomes the sink.

If the region enclosed by the polygon of Fig. 44 is now mapped upon the upper half of a z-plane in such a way that the origin and the point at infinity for this plane are identified respectively with the vertexes A and B of the polygon, the equivalent flow map in the z-plane is simply given by the concentric circles about the origin (equipotential loci) and the radial lines extending from the origin to infinity (flow lines). The vertex at the point C in Fig. 44 is then represented by a finite point on the real axis of the z-plane.

In view of the fact that the points on the real axis of the z-plane corresponding to the vertexes of the polygon must be arranged in such an order that they are encountered in the same sequence during a traversal

of the real axis in the positive direction (leaving the upper half plane to the left) as are the vertexes of the polygon when its boundary is traversed in the corresponding positive direction (leaving the enclosed region to the left), it follows that the point on the real axis of the z-plane corresponding to the vertex C must lie to the left of the origin as indicated in Fig. 45. This is the point designated as $z = z_1$. Since this vertex is like the one shown in part (c) of Fig. 40, the corresponding exponent μ_1 has the value -1. The vertex A, which corresponds to $z = 0$, is like the one shown in Fig. 41. The μ-value for this vertex, therefore, is $+1$. As pointed out in the preceding discussion, a vertex which corresponds to the point $z = \infty$ is not represented by a factor of the form $(\zeta - z_\nu)^{-\mu_\nu}$ in the integral 547 but comes about by virtue of the behavior of this function in the vicinity of the point at infinity. In other words, the corresponding factor is implicitly rather than explicitly contained in the general form of the integral, which for the present problem is now recognized to read

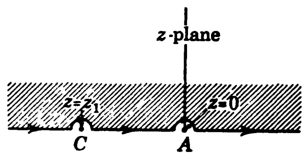


FIG. 45. Contour in the z-plane for the problem of Fig. 44.

$$w = M \int_{z_0}^{z} \frac{d\zeta}{(\zeta - z_1)^{-1}(\zeta - 0)^1} + N \tag{581}$$

Thus for the vertex B (corresponding to $z = \infty$) one has, according to Eq. 574,

$$\Delta\theta = (2 + 1 - 1)\pi = 2\pi \tag{582}$$

which checks with the above statement that this vertex has the character of w_ν in Fig. 43 for $\mu_\nu = 2$.

Equation 581 may more appropriately be written in the form

$$w = M \int_{z_0}^{z} \frac{(\zeta - z_1)\, d\zeta}{\zeta} + N \tag{583}$$

It is now convenient to choose the lower limit z_0 equal to z_1, corresponding to the vertex C in Fig. 44. Here w must have the value jd, and since the integral in Eq. 583 vanishes for $z = z_0 = z_1$, it follows that

$$N = jd \tag{584}$$

The constant M may be evaluated through calculating the increment in the function w corresponding to the increment in z resulting from a traversal of the semicircular detour about the point $z = 0$. For this semicircular path one may write

$$\zeta = \rho e^{j\phi} \tag{585}$$

in which ρ is the radius of the small arc. For small values of ζ, the factor $(\zeta - z_1)$ in the integral of Eq. 583 may be replaced by $-z_1$, and inasmuch as Eq. 585 yields

$$\frac{d\zeta}{\zeta} = j\,d\phi \qquad [586]$$

it is seen that the increment in the function w corresponding to the arc increment in z becomes

$$\Delta w = -z_1 Mj \int_{\pi}^{0} d\varphi = j\pi z_1 M \qquad [587]$$

This increment in w must equal the change in the value of w which corresponds to passing around the vertex A. With reference to Fig. 44 one sees that this change equals the increment in w between the bottom face of the condenser plate and the center line, and hence that this increment in w must equal $-jd$. Thus

$$\Delta w = -jd = j\pi z_1 M \qquad [588]$$

whence

$$M = -\frac{d}{\pi z_1} \qquad [589]$$

The representation for the function w now reads

$$w = -\frac{d}{\pi z_1}\int_{z_1}^{z} \frac{(\zeta - z_1)\,d\zeta}{\zeta} + jd \qquad [590]$$

The integration, which is straightforward, yields

$$w = \frac{d}{\pi}\left[1 - \frac{z}{z_1} + \ln z - \ln z_1\right] + jd \qquad [591]$$

Since there are no further conditions to be satisfied, it appears that the location of the point z_1 on the negative real axis in Fig. 45 may be chosen arbitrarily. A choice which yields a simple form for the resulting mapping function is $z_1 = -1$. Since $\ln(-1) = j\pi$, Eq. 591 becomes

$$w = \frac{d}{\pi}[1 + z + \ln z] \qquad [592]$$

This is the desired mapping function which converts the family of concentric circles about the origin of the z-plane and the orthogonal family of straight lines radiating from the origin (both confined to the upper half plane) into the orthogonal families of loci which map the field and the equipotential lines for the condenser plate in Fig. 44. If desired, this map may be converted to the rectangular reference form shown in the

t-plane of Fig. 46 through making the further change of variable given by Eq. 544. Then Eq. 592 becomes

$$w = \frac{d}{\pi}[1 + t + e^t] \tag{593}$$

The reference field in the t-plane, with several points of particular interest marked on it, is shown in Fig. 46. The edge C of the condenser plate in Fig. 44 where $w = jd$ corresponds to $t = j\pi$. The under side of the condenser plate in Fig. 44 corresponds to the portion of the horizontal line $n = j\pi$ to the left of C; the top side of the plate corresponds to the portion of this line to the right of C. The center line in Fig. 44 is given by the real axis of the t-plane. The transformation 593 is seen virtually to remove the 360-degree bend in the boundary at the edge C in Fig. 44 so that the top side of the condenser plate appears as a linear extension of the bottom side. It may also be of interest to note that the origin of Fig. 44 does not appear directly below the point C in the t-plane but is shifted to the negative real point $t = -1.279$. The process of transferring the rectangular grid of lines in the t-plane of Fig. 46 to the region of interest in Fig. 44 by means of the function 593 is left as an exercise for the reader.

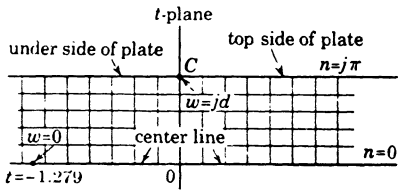


FIG. 46. The field of the parallel-plate condenser in the t-plane.

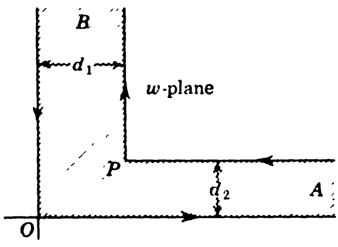


FIG. 47. Relevant to plotting the flow in a right-angle bend.

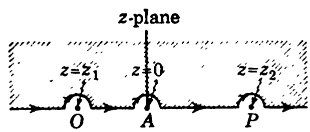


FIG. 48. The contour in the z-plane for the problem of Fig. 47.

As a second example let it be required to determine the field map for the flow around the right-angle bend indicated by the shaded region in the w-plane of Fig. 47. The polygon in this example has four vertexes. One lies at the origin O, another at the finite point P, and the remaining two at infinity. Both these have the character of the vertex w_r of Fig. 42.

They are here denoted as the vertexes A and B. They become the source and sink respectively for the present problem.

The points on the real axis of the z-plane which are chosen to correspond to these vertexes are indicated in Fig. 48. The source vertex A is assumed to lie at the origin and the sink vertex B at infinity. The proper order, as determined from the corresponding directions of traversal, requires that the vertexes O and P lie respectively to the left and to the right of the origin. These are designated as the points $z = z_1$ and $z = z_2$.

Recognizing that the exponents μ_ν for the vertexes O, A, P are respectively $+\frac{1}{2}$, $+1$, and $-\frac{1}{2}$, and choosing the lower limit z_0 of the integral 547 equal to z_1, one has

$$w = M \int_{z_1}^{z} \frac{d\zeta}{(\zeta - z_1)^{1/2}(\zeta - 0)^1(\zeta - z_2)^{-1/2}} + N \qquad [594]$$

or

$$w = M \int_{z_1}^{z} \frac{(\zeta - z_2)^{1/2}\, d\zeta}{\zeta(\zeta - z_1)^{1/2}} + N \qquad [595]$$

Since the origin of the w-plane corresponds to the point $z = z_1$, one has $w(z_1) = 0$, and hence Eq. 595 yields

$$N = 0 \qquad [596]$$

The reason for choosing the lower limit of the integral as the point $z = z_1$ is thus evident.

In order to study the function in the vicinity of the vertex B which occurs for $z = \infty$, it is necessary to make the change of variable

$$\zeta^* = \frac{-1}{\zeta} \qquad [597]$$

whence

$$\frac{d\zeta}{\zeta} = -\frac{d\zeta^*}{\zeta^*} \qquad [598]$$

The vicinity $\zeta = \infty$ is identified with the vicinity $\zeta^* = 0$. The increment in the function w resulting from passing around the vertex B at $\zeta^* = 0$ is determined through considering the integral 595 for $\zeta \to \infty$ and making the substitutions 597 and 598. This gives

$$\Delta w = -M \int_{-\rho}^{\rho} \frac{d\zeta^*}{\zeta^*} \qquad [599]$$

in which the integration is to be extended over a small semicircular arc of radius ρ about the origin of the ζ^*-plane. For this arc

$$\zeta^* = \rho e^{j\phi} \qquad [600]$$

and

$$\frac{d\zeta^*}{\zeta^*} = j\,d\phi \qquad [601]$$

so that the integral 599 becomes

$$\Delta w = -jM\int_\pi^0 d\phi = j\pi M \qquad [602]$$

But " passing around " the vertex B in the w-plane amounts to going from the right- to the left-hand boundary of the cross-hatched region B in Fig. 47 and hence yields an increment in the function w equal to $-d_1$. Therefore

$$\Delta w = -d_1 = j\pi M \qquad [603]$$

In a similar manner, passing around the vertex A corresponding to the vicinity of $\zeta = 0$ yields

$$\Delta w = M\sqrt{\frac{z_2}{z_1}}\int_\rho^\rho \frac{d\zeta}{\zeta} = jM\sqrt{\frac{z_2}{z_1}}\int_\pi^0 d\phi = -j\pi M\sqrt{\frac{z_2}{z_1}} \qquad [604]$$

But according to Fig. 47 this must equal jd_2. Hence

$$jd_2 = -j\pi M\sqrt{\frac{z_2}{z_1}} \qquad [605]$$

Equations 603 and 605 determine the constant M and the ratio z_2/z_1. Since there are no further conditions to be fulfilled, one of the points z_1 or z_2 may be chosen arbitrarily. For the moment this choice is left open. The ratio is found to be given by

$$\sqrt{\frac{z_2}{z_1}} = j\frac{d_2}{d_1} \quad \text{or} \quad \frac{z_2}{z_1} = -\frac{d_2^2}{d_1^2} \qquad [606]$$

This result agrees with the fact that z_2 is positive and z_1 negative.

The required function is now represented by

$$w = \frac{jd_1}{\pi}\int_{z_1}^{z} \frac{\sqrt{\zeta - z_2}}{\zeta\sqrt{\zeta - z_1}}\,d\zeta \qquad [607]$$

which may be converted into

$$w = \frac{jd_1}{\pi}\int_{z_1}^{z} \frac{\zeta - z_2}{\zeta\sqrt{(\zeta - z_1)(\zeta - z_2)}}\,d\zeta \qquad [608]$$

or

$$w = \frac{jd_1}{\pi}\int_{z_1}^{z} \frac{d\zeta}{\sqrt{(\zeta - z_1)(\zeta - z_2)}} - \frac{jd_1z_2}{\pi}\int_{z_1}^{z} \frac{d\zeta}{\zeta\sqrt{(\zeta - z_1)(\zeta - z_2)}} \qquad [609]$$

If in the second integral one makes the change of variable indicated by Eq. 597 and also the consistent changes

$$z^* = -\frac{1}{z}, \quad z^*_{1,2} = \frac{-1}{z_{1,2}} \tag{610}$$

with the help of Eq. 606, one finds

$$w = \frac{jd_1}{\pi}\int_{z_1}^{z} \frac{d\zeta}{\sqrt{(\zeta - z_1)(\zeta - z_2)}} + \frac{d_2}{\pi}\int_{z^*_1}^{z^*} \frac{d\zeta^*}{\sqrt{(\zeta^* - z^*_1)(\zeta^* - z^*_2)}} \tag{611}$$

The integration yields

$$w = \frac{jd_1}{\pi}\ln\left\{\frac{z - \frac{1}{2}(z_1 + z_2) + \sqrt{(z - z_1)(z - z_2)}}{\frac{1}{2}(z_1 - z_2)}\right\} + \frac{d_2}{\pi}\ln\left\{\frac{z^* - \frac{1}{2}(z^*_1 + z^*_2) + \sqrt{(z^* - z^*_1)(z^* - z^*_2)}}{\frac{1}{2}(z^*_1 - z^*_2)}\right\} \tag{612}$$

Inasmuch as an arbitrary specification with regard to z_1 or z_2 still remains open, it is possible, in agreement with Eq. 606, to set

$$z_2 = \frac{d_1}{d_2} \quad \text{and} \quad z_1 = -\frac{d_2}{d_1} \tag{613}$$

Then Eq. 612 takes the form

$$w = d_1 + \frac{jd_1}{\pi}f(z) + \frac{d_2}{\pi}f(z^*) \tag{614}$$

in which

$$f(z) = \ln\left\{\frac{z + \dfrac{d_2^2 - d_1^2}{2d_1d_2} + \sqrt{\left(z + \dfrac{d_2}{d_1}\right)\left(z - \dfrac{d_1}{d_2}\right)}}{\dfrac{d_1^2 + d_2^2}{2d_1d_2}}\right\} \tag{615}$$

For purposes of calculation it may be helpful to note that the substitution

$$z' = \frac{z + \dfrac{1}{2}\left(\dfrac{d_2}{d_1} - \dfrac{d_1}{d_2}\right)}{\dfrac{1}{2}\left(\dfrac{d_2}{d_1} + \dfrac{d_1}{d_2}\right)} \tag{616}$$

converts the expression 615 to the form

$$f(z') = \ln\left(z' + \sqrt{(z')^2 - 1}\right) \tag{617}$$

If $d_1 = d_2$, the resulting mapping function is greatly simplified inasmuch as Eq. 616 shows that then $z' = z$.

..ne previous example, the field map may be reduced to points ..ar grid by means of the additional substitution $z = e^t$.

26. Hurwitz polynomials; stability criteria

In problems dealing with the dynamics of a physical system one is frequently concerned with the question of the stability of its behavior. In terms of the so-called characteristic equation of the system, which has the form

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0 \qquad [618]$$

with real coefficients $a_0 \cdots a_n$, this question is answered in the affirmative if it can be established that all the real roots and all the real parts of the complex roots are negative. Stated in another way, the stability of the physical system is assured if it can be shown that all the zeros of the polynomial

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \qquad [619]$$

lie in the left half of the z-plane. A polynomial having this property is called a *Hurwitz polynomial.*

The necessary and sufficient conditions which the coefficients of an arbitrary polynomial (with real coefficients) must satisfy in order that it be a Hurwitz polynomial are spoken of as the *Hurwitz criteria.* In dynamics these same conditions are alternatively referred to as *Routh's stability criteria.*

Starting with the factored form of the polynomial 619, which reads

$$P(z) = a_n (z - z_1)(z - z_2) \cdots (z - z_n) \qquad [620]$$

one may readily establish (by multiplying out and collecting terms with like powers of z) that

$$\begin{aligned} \frac{a_{n-1}}{a_n} &= -(z_1 + z_2 + \cdots + z_n) \\ \frac{a_{n-2}}{a_n} &= z_1 z_2 + z_1 z_3 + \cdots + z_{n-1} z_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{a_0}{a_n} &= (-1)^n \cdot z_1 z_2 \cdots z_n \end{aligned} \qquad [621]$$

If it is assumed that all the roots are real and negative, and that $a_n > 0$, it is evident that all the coefficients are positive. If some or all of the roots are in the form of conjugate complex pairs, then one may likewise establish (by the use of Eqs. 621) that all the coefficients are positive if the

roots have negative real parts. It is thus seen that the positiveness of the coefficients* is a necessary condition for the negativeness of the real parts of the roots of a given algebraic equation, but this condition alone proves to be insufficient, as is shown by the following discussion.

One begins by investigating the significant properties of Hurwitz polynomials. In these considerations two simple facts form the basis from which the pertinent conclusions are readily obtained. These facts are, first, that the zeros of the polynomial $P(z)$ lie in the left half of the z-plane and, second, that their distribution in this half plane is symmetrical about the real axis. The last statement is evident from the fact that complex roots must occur in conjugate pairs.

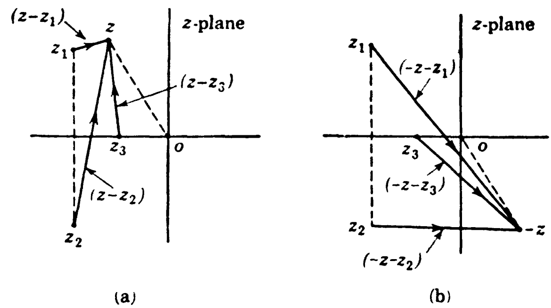


FIG. 49. Relevant to the properties of a third degree Hurwitz polynomial.

The polynomial is considered to be in its factored form as given by Eq. 620. For an arbitrary value of the complex variable z, the several factors for a polynomial of the third degree may be represented graphically as shown in Fig. 49. Parts (a) and (b) of this figure show how the graphical representation of the factors changes when the algebraic sign of the variable z is reversed. It is clear that if z is chosen to lie in the left half plane (as in part (a) of Fig. 49) $-z$ lies in the right half plane, and vice versa. It may also be seen that if z is replaced by its conjugate value $\bar{z}$, the factors collectively represent the same set of *magnitudes* since pairs of factors representing conjugate roots merely become interchanged. This is, of course, also obvious from the fact that $P(\bar{z})$ is the conjugate of $P(z)$, and hence these two values of the polynomial have the same magnitude. On the other hand, if the point z is replaced by its image about the imaginary axis (this amounts to replacing z by its negative conjugate value), collectively the magnitudes of the factors change in the same way as they do when z is replaced by $-z$. Points which are images

*It is obvious that one may alternatively state that all the coefficients must be negative, the significant point being that they all have the same algebraic sign.

with respect to the imaginary axis are regarded as corresponding points in the left and right half planes.

It should be clear from the representations in Fig. 49 that for any point z in the right half plane, the magnitude of the polynomial is larger than it is for the corresponding point in the left half plane. Together with the considerations of the preceding paragraph, this fact enables one to see without difficulty the truth of the following statements:

$$\begin{aligned} |P(z)| &> |P(-z)| \quad \text{for } Re(z) > 0 \\ |P(z)| &= |P(-z)| \quad \text{for } Re(z) = 0 \\ |P(z)| &< |P(-z)| \quad \text{for } Re(z) < 0 \end{aligned} \tag{622}$$

in which Re denotes " real part of." Letting

$$\phi(z) = \frac{P(z)}{P(-z)} \tag{623}$$

one may alternatively express these results by

$$\begin{aligned} |\phi(z)| &> 1 \quad \text{for } Re(z) > 0 \\ |\phi(z)| &= 1 \quad \text{for } Re(z) = 0 \\ |\phi(z)| &< 1 \quad \text{for } Re(z) < 0 \end{aligned} \tag{624}$$

It should be clearly recognized that these statements hold only if $P(z)$ is a Hurwitz polynomial, for if $P(z)$ has any zeros in the right half plane, points can certainly be found for which these statements collectively are no longer true. One may, therefore, conclude that if the conditions 622 or 624 are true, $P(z)$ must be a Hurwitz polynomial. The relations 622 or 624 are the necessary and sufficient conditions that a given polynomial $P(z)$ have zeros in the left half plane only. By means of the succeeding manipulations, these conditions are put into a more usable form.

The first step in this direction is to introduce the function

$$\psi(z) = \frac{\phi(z) + 1}{\phi(z) - 1} = \frac{P(z) + P(-z)}{P(z) - P(-z)} \tag{625}$$

According to the discussion given in Art. 24, this transformation maps the interior of the unit circle in the ϕ-plane upon the left half of the ψ-plane. Hence one has

$$\begin{aligned} Re(\psi) &> 0 \quad \text{for } |\phi(z)| > 1 \\ Re(\psi) &= 0 \quad \text{for } |\phi(z)| = 1 \\ Re(\psi) &< 0 \quad \text{for } |\phi(z)| < 1 \end{aligned} \tag{626}$$

and, with the use of the relations 624, one obtains

$$\begin{aligned} Re(\psi) &> 0 \quad \text{for } Re(z) > 0 \\ Re(\psi) &= 0 \quad \text{for } Re(z) = 0 \\ Re(\psi) &< 0 \quad \text{for } Re(z) < 0 \end{aligned} \qquad [627]$$

If the polynomial $P(z)$ is written in the form

$$P(z) = m(z) + n(z) \qquad [628]$$

in which $m(z)$ represents the terms involving even powers of z (called the even part of P) and $n(z)$ represents the terms involving odd powers of z (called the odd part of P), according to Eq. 625, one has

$$\psi(z) = \frac{m(z)}{n(z)} \qquad [629]$$

It is thus established that if $m(z)$ and $n(z)$ are the even and odd parts respectively of a Hurwitz polynomial, the rational function 629 has the properties expressed by the conditions 627, and, conversely, if the ratio of the even and odd parts of a given polynomial yields a rational function having the properties 627, that polynomial must be a Hurwitz polynomial.

These properties are now examined in greater detail. Suppose the rational function $\psi(z)$ has a pole of the order s at some point $z = z_\nu$. The Laurent series for $\psi(z)$ in this vicinity then reads

$$\psi(z) = \frac{b_{-s}}{(z - z_\nu)^s} + \cdots + \frac{b_{-1}}{(z - z_\nu)^1} + b_0 + b_1(z - z_\nu) + \cdots \qquad [630]$$

For points very close to z_ν one may write

$$\psi(z) \cong \frac{b_{-s}}{(z - z_\nu)^s} \qquad [631]$$

Letting

$$b_{-s} = ke^{j\beta} \qquad [632]$$

and

$$(z - z_\nu) = \rho e^{j\alpha} \qquad [633]$$

one has

$$\psi(z) \cong \frac{k}{\rho^s} e^{j(\beta - s\alpha)} \qquad [634]$$

whence

$$Re(\psi) \cong \frac{k}{\rho^s} \cos(s\alpha - \beta) \qquad [635]$$

It is thus seen that, in the immediate vicinity of the pole, the real part

of $\psi(z)$ assumes large negative as well as large positive values. More specifically, as α is allowed to vary from 0 to 2π (the vicinity of the pole is explored through passing around it on a concentric circle of small radius), the real part of $\psi(z)$ is observed to change sign $2s$ times.

With reference to the conditions 627, one is forced to conclude immediately that the function $\psi(z)$ cannot have poles in either the right or the left half plane. This fact restricts the poles of $\psi(z)$ to lie along the imaginary axis of the z-plane. But the conditions 627 together with the Eq. 635 impose further restrictions upon these poles. The only conditions under which Eq. 635 for a pole on the imaginary axis does not conflict with the restrictions 627 are that $\beta = 0$ and $s = 1$, for then Eq. 635 reads

$$Re(\psi) \cong \frac{k}{\rho} \cos \alpha \qquad [636]$$

According to Eq. 633, $Re(z) > 0$ corresponds to $-\pi/2 < \alpha < \pi/2$, and $Re(z) < 0$ corresponds to $\pi/2 < \alpha < 3\pi/2$, whereas for $Re(z) = 0$, $\alpha = \pm\pi/2$. Equation 636 is thus seen to yield a real part of $\psi(z)$, which behaves in agreement with the conditions 627. The restriction that $s = 1$ means that the pole must be simple, and $\beta = 0$ requires that the residue of $\psi(z)$ in this pole be real and positive.

The conditions 627, therefore, require that the rational function $\psi(z)$ have poles on the imaginary axis only and that these poles be simple and have positive real residues. In order to see that these requirements on the function $\psi(z)$ are also sufficient to assure the fulfillment of the conditions 627, one need merely regard a typical term in the partial fraction expansion of $\psi(z)$. Such a term reads

$$\frac{k_\nu}{(z - z_\nu)} \qquad [637]$$

Since the residue k_ν is real and positive, and z_ν is a pure imaginary quantity, it is evident that the term 637 has the properties demanded by the conditions 627, and hence the finite sum of such terms which represents $\psi(z)$ has these properties.

One has thus gained a new formulation for the necessary and sufficient conditions that $P(z)$ be a Hurwitz polynomial. Namely, the quotient of its even and odd parts must be a function having simple poles on the imaginary axis only, and with positive real residues in these poles.

It is collaterally useful to digress for a moment and study somewhat more carefully the properties of the function $\psi(z)$. First it should be observed that if $\psi(z)$ satisfies the conditions 627, its reciprocal $1/\psi(z)$ does so also. Hence, using Eq. 629, the function

$$\frac{1}{\psi(z)} = \frac{n(z)}{m(z)} \qquad [638]$$

must also have simple poles on the imaginary axis only and positive real residues. Both the even and odd parts $m(z)$ and $n(z)$ must, therefore, be polynomials whose zeros are simple and lie on the imaginary axis. Furthermore, since poles of $\psi(z)$ or $1/\psi(z)$ which may lie at $z = 0$ or at $z = \infty$ must also be simple, it follows that the highest powers of $m(z)$ and $n(z)$ as well as their lowest powers may not differ by more than unity. They obviously must differ at least by unity because $m(z)$ and $n(z)$ are respectively even and odd.

If one writes

$$\psi(z) = u(x,y) + jv(x,y) \tag{639}$$

the conditions 627 are seen to yield

$$\frac{\partial u}{\partial x} > 0 \quad \text{for } x = 0 \tag{640}$$

The Cauchy-Riemann Eq. 12 then shows that

$$\frac{\partial v}{\partial y} > 0 \quad \text{for } x = 0 \tag{641}$$

But by Eqs. 627, $u = 0$ for $x = 0$, so that this result may alternatively be written

$$\frac{\partial \psi}{j\,\partial y} > 0 \quad \text{for } x = 0 \tag{642}$$

which states that along the imaginary axis (where ψ has pure imaginary values only) ψ is a continuously increasing function. The interesting consequence of this fact is that the zeros and poles of $\psi(z)$, which lie along the imaginary axis, must alternate.

If the polynomials $m(z)$ and $n(z)$ are written in their factored forms, the expression 629 for $\psi(z)$ reads

$$\psi(z) = \frac{a_n(z^2 - z_1{}^2)(z^2 - z_3{}^2)\cdots(z^2 - z^2{}_{2n-1})}{a_{n-1}z(z^2 - z_2{}^2)(z^2 - z_4{}^2)\cdots(z^2 - z^2{}_{2n-2})} \tag{643}$$

in which it is arbitrarily assumed that the polynomial $P(z)$ is of the even degree $2n$. The alternation of zeros and poles along the imaginary axis is expressed by the conditions

$$0 < |z_1| < |z_2| < \cdots < |z_{2n-2}| < |z_{2n-1}| < \infty \tag{644}$$

This result is referred to as the *separation property* of the zeros and poles of $\psi(z)$.

Using Eq. 220 for the evaluation of the residues of $\psi(z)$ in any one of

its poles at $z = z_2, z_4, \cdots z_{2n-2}$, one has

$$k_\nu = [(z - z_\nu)\psi(z)]_{z=z_\nu}$$
$$= \frac{a_n(z_\nu^2 - z_1^2)(z_\nu^2 - z_3^2)\cdots(z_\nu^2 - z^2_{2n-1})}{a_{n-1}2z_\nu^2(z_\nu^2-z_2^2)\cdots(z_\nu^2-z^2_{\nu-2})(z_\nu^2-z^2_{\nu+2})\cdots(z_\nu^2-z^2_{2n-2})} \quad [645]$$

On the assumption that a_n/a_{n-1} is positive, it may be seen from this result that the separation property expressed by the inequalities 644 assures the positiveness of the residues of $\psi(z)$ in all its poles at finite frequencies. The residue of $\psi(z)$ at $z = \infty$, incidentally, is recognized to be a_n/a_{n-1}, whereas the one at $z = 0$ has the value of $z\psi(z)$ for $z = 0$, which is positive since all the quantities $-z_\nu^2$ are positive.

In view of these further detailed results it may be stated that if $P(z)$ is a polynomial with positive coefficients, and if its even and odd parts $m(z)$ and $n(z)$ differ in their highest and in their lowest powers by no more than unity and have simple zeros restricted to the imaginary axis where they mutually separate each other, $P(z)$ is a Hurwitz polynomial.

The purpose in thus stating in a variety of forms, the necessary and sufficient conditions that $P(z)$ be a Hurwitz polynomial is to focus attention upon properties of these polynomials which frequently become collaterally useful. The actual process of making such a test may be based directly upon any one of the sets of conditions already stated. A particularly effective procedure, however, is derived from these considerations in the following manner.

For any given polynomial, the function $\psi(z)$ is readily formed according to Eq. 629. If $P(z)$ is of an even degree, $\psi(z)$ has a pole at infinity; otherwise $1/\psi(z)$ has a pole at infinity. Whichever may be the case, one begins the procedure by considering that function (ψ or $1/\psi$) which does have a pole at infinity. Without any loss to the ensuing argument, this function is assumed to be $\psi(z)$. More completely represented, it has the form

$$\psi(z) = \frac{a_nz^n + a_{n-2}z^{n-2} + \cdots + a_2z^2 + a_0}{a_{n-1}z^{n-1} + a_{n-3}z^{n-3} + \cdots + a_1z} \quad [646]$$

The first step in the procedure is to divide the denominator polynomial into the numerator polynomial by the common process of long division, however, ceasing after only a single term in the quotient is determined. This yields

$$\psi(z) = \frac{a_n}{a_{n-1}}z + \frac{a'_{n-2}z^{n-2} + a'_{n-4}z^{n-4} + \cdots + a'_2z^2 + a'_0}{a_{n-1}z^{n-1} + a_{n-3}z^{n-3} + \cdots + a_1z} \quad [647]$$

which may be indicated more compactly as

$$\psi(z) = \frac{a_n}{a_{n-1}}z + \frac{m'(z)}{n(z)} \quad [648]$$

If $P(z)$, from which $\psi(z)$ is derived, is a Hurwitz polynomial, $\psi(z)$ must have simple poles on the imaginary axis with positive residues. In particular the pole at infinity, represented by the first term in Eq. 648, must yield a positive residue. Hence one has the first particular condition

$$\frac{a_n}{a_{n-1}} > 0 \tag{649}$$

If $\psi(z)$ is imagined to be represented by its partial fraction expansion, one may identify the first term in Eq. 648 with that term in this expansion which represents the pole at infinity. The remaining terms in this partial fraction expansion are then seen to represent the corresponding expansion for the function given by the second term in Eq. 648. This function is the remainder after the pole at infinity is removed from $\psi(z)$. It thus becomes clear that this remainder function must have the same properties as $\psi(z)$, namely, simple poles restricted to the imaginary axis and positive real residues, and its reciprocal must likewise have these properties.

Denoting the reciprocal of the remainder function by $\psi^*(z)$, one has, according to Eqs. 647 and 648,

$$\psi^*(z) = \frac{a_{n-1}z^{n-1} + a_{n-3}z^{n-3} + \cdots + a_1 z}{a'_{n-2}z^{n-2} + a'_{n-4}z^{n-4} + \cdots + a'_2 z^2 + a'_0} \tag{650}$$

This function evidently again has a simple pole at infinity. Repeating the process of long division as before, one obtains

$$\psi^*(z) = \frac{a_{n-1}}{a'_{n-2}}z + \frac{a'_{n-3}z^{n-3} + a'_{n-5}z^{n-5} + \cdots + a'_1 z}{a'_{n-2}z^{n-2} + a'_{n-4}z^{n-4} + \cdots + a'_2 z^2 + a'_0} \tag{651}$$

or more compactly

$$\psi^*(z) = \frac{a_{n-1}}{a'_{n-2}}z + \frac{n'(z)}{m'(z)} \tag{652}$$

The positiveness of the residue of $\psi^*(z)$ in its pole at infinity requires the second particular condition

$$\frac{a_{n-1}}{a'_{n-2}} > 0 \tag{653}$$

The second term in Eq. 652 is a subsequent remainder function which again must have the same properties as $\psi(z)$, and its reciprocal must also have these properties. The inverted remainder function

$$\psi^{**}(z) = \frac{m'(z)}{n'(z)} \tag{654}$$

again has a simple pole at infinity with a residue which must be positive.

The continuation of the process is thus clear, and leads successively to additional particular conditions like the ones expressed by the inequalities 649 and 653. The procedure terminates after all the terms in $m(z)$ and $n(z)$ are exhausted.

Letting

$$\begin{aligned} \alpha_1 &= \frac{a_n}{a_{n-1}} \\ \alpha_2 &= \frac{a_{n-1}}{a'_{n-2}} \\ \alpha_3 &= \frac{a'_{n-2}}{a'_{n-3}} \\ &\cdots\cdots\cdots \end{aligned} \tag{655}$$

one obtains finally a representation for $\psi(z)$ of the form

$$\psi(z) = \alpha_1 z + \cfrac{1}{\alpha_2 z + \cfrac{1}{\alpha_3 z + \cfrac{\cdot}{\cdot \; + \cfrac{1}{\alpha_n z}}}} \tag{656}$$

which is referred to as a finite *Stieltjes* continued fraction.* It contains altogether n terms, as is clear from the following series of fractions indicating merely the degrees of the numerator and denominator polynomials appearing in the original function $\psi(z)$ and in the successively encountered inverted remainder functions $\psi^*(z)$. $\psi^{**}(z)$, etc.

$$\frac{n}{n-1}, \quad \frac{n-1}{n-2}, \quad \frac{n-2}{n-3}, \quad \cdots \quad \frac{2}{1}, \quad \frac{1}{0} \tag{657}$$

The necessary and sufficient conditions that $P(z)$ be a Hurwitz polynomial are now simply expressed by the statement that all the quantities α_1, α_2, α_3, etc., as given by Eqs. 655 must be positive.

This procedure for testing a given polynomial may be replaced by a complementary one in which terms representing poles at $z = 0$ are removed from $\psi(z)$ and the successive inverted remainder functions. One may say that this variation in the method amounts merely to replacing z by $1/z$ and proceeding as discussed above. The test is thus applied to the polynomial $P(1/z)$ instead of to $P(z)$. Since the transforma-

*This is one having the form of Eq. 656 in which all the coefficients are positive real numbers.

tion $z \to 1/z$ maps the left half plane upon itself, it is clear that $P(z)$ is proved to be a Hurwitz polynomial if it can be shown that $P(1/z)$ is such a one.

To carry out this variation in the procedure one begins again with the expression 646 for $\psi(z)$ but with the polynomials turned end-for-end. The first step in the process of dividing the denominator into the numerator yields

$$\psi(z) = \frac{a_0}{a_1 z} + \frac{a''_2 z^2 + \cdots + a''_n z^n}{a_1 z + \cdots + a_{n-1} z^{n-1}} \qquad [658]$$

The first particular condition reads

$$\frac{a_0}{a_1} > 0 \qquad [659]$$

The inverted remainder must again have the same properties as $\psi(z)$ because the second term in Eq. 658 is equal to the partial fraction expansion of $\psi(z)$ minus the term for the pole at $z = 0$. This second term, therefore, has the stated properties and so does its reciprocal. The inverted remainder, moreover, again has a simple pole at $z = 0$, and the requirement that the residue in this pole be positive yields the second particular condition

$$\frac{a_1}{a''_2} > 0 \qquad [660]$$

and so forth.

One thus arrives at the alternate finite Stieltjes continued fraction

$$\psi(z) = \beta_1 z^{-1} + \cfrac{1}{\beta_2 z^{-1} + \cfrac{\ddots}{\quad + \cfrac{1}{\beta_n z^{-1}}}} \qquad [661]$$

in which

$$\beta_1 = \frac{a_0}{a_1}$$

$$\beta_2 = \frac{a_1}{a''_2} \qquad [662]$$

$$\beta_3 = \frac{a''_2}{a''_3}$$

$$\cdots\cdots$$

The necessary and sufficient conditions that $P(z)$ be a Hurwitz polynomial are that all the coefficients $\beta_1 \cdots \beta_n$ be positive. These conditions are entirely equivalent to the ones involving the coefficients $\alpha_1 \cdots \alpha_n$ given by Eqs. 655, although they are different in their detailed appearance. It should be clear, however, that the quantities $\beta_1 \cdots \beta_n$ as functions of the coefficients $a_0, a_1, \cdots a_n$ of the given polynomial $P(z)$ must be the same as the expressions for $\alpha_1 \cdots \alpha_n$ in terms of these coefficients except that the consecutive order of the subscripts 0, 1, 2, $\cdots$ n on the a's is inverted. In other words, if in the expressions for $\alpha_1 \cdots \alpha_n$ one makes the substitutions indicated by

$$\begin{array}{l} a_n \quad \rightarrow a_0 \\ a_{n-1} \rightarrow a_1 \\ \cdots\cdots\cdots\cdots \\ a_1 \quad \rightarrow a_{n-1} \\ a_0 \quad \rightarrow a_n \end{array} \qquad [663]$$

the corresponding ones for $\beta_1 \cdots \beta_n$ are obtained.

In order to determine the expressions for the α's or β's in terms of the coefficients $a_0 \cdots a_n$ of the polynomial $P(z)$ from the results expressed by Eqs. 655 and 662, it is necessary to carry through the indicated processes of long division involved in the steps leading to the continued fraction expansions 656 and 661, so as to obtain the coefficients $a'_0 \cdots a'_{n-2}, a''_2 \cdots a''_{n-4}$, etc., as functions of $a_0 \cdots a_n$. If this is done for the β's, one finds that the results obtained for the first three steps (that is, for the determination of $\beta_1, \beta_2, \beta_3$) yield conditions which may be expressed in the form

$$\begin{aligned} D_1 &= a_1 > 0 \\ D_2 &= \begin{vmatrix} a_1 & a_0 \\ a_3 & a_2 \end{vmatrix} > 0 \\ D_3 &= \begin{vmatrix} a_1 & a_0 & 0 \\ a_3 & a_2 & a_1 \\ a_5 & a_4 & a_3 \end{vmatrix} > 0 \end{aligned} \qquad [664]$$

with the condition $a_0 > 0$ tacitly understood. From the structure of these determinants and the recurrent nature of the process of derivation, one may assume that the determinant of nth order has the form

$$D_n = \begin{vmatrix} a_1 & a_0 & 0 & & & \cdots & & 0 \\ a_3 & a_2 & a_1 & a_0 & 0 & \cdots & & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & a_0 & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{2n-1} & a_{2n-2} & & & \cdots & & & a_n \end{vmatrix} \qquad [665]$$

The correctness of this generalization may be verified through showing that it is consistent with the process of deriving an nth degree Hurwitz polynomial from one of the degree $n - 1$. Suppose the latter is written in the form

$$P_{n-1} = a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + a_n \qquad [666]$$

in which the order of the coefficients is reversed, so that the expressions for β in the former polynomial become identical with those for α in the present one. The corresponding function $\psi(z)$ reads

$$\psi_{n-1} = \frac{a_2 z^{n-2} + a_4 z^{n-4} + \cdots + a_{n-2} z^2 + a_n}{a_1 z^{n-1} + a_3 z^{n-3} + \cdots + a_{n-1} z} \qquad [667]$$

Since this function has a zero at infinity, the one associated with a Hurwitz polynomial of the nth degree may be formed by adding to the expression 667 a term representing a simple pole at infinity with a positive residue. This function may, therefore, be written

$$\psi_n = \psi_{n-1} + a_0 z \qquad [668]$$

with the condition that $a_0 > 0$. Using Eq. 667, one finds

$$\psi_n = \frac{a_0 a_1 z^n + (a_0 a_3 + a_2) z^{n-2} + (a_0 a_5 + a_4) z^{n-4} + \cdots + (a_0 a_{n-1} + a_{n-2}) z^2 + a_n}{a_1 z^{n-1} + a_3 z^{n-3} + \cdots + a_{n-1} z} \qquad [669]$$

and the associated polynomial reads

$$\begin{aligned} P_n = a_0 a_1 z^n &+ a_1 z^{n-1} + (a_0 a_3 + a_2) z^{n-2} + a_3 z^{n-3} \\ &+ (a_0 a_5 + a_4) z^{n-4} + \cdots + a_{n-1} z + a_n \end{aligned} \qquad [670]$$

From the manner of its derivation, P_n is a Hurwitz polynomial if P_{n-1} is known to be such a one, and if a_0 is positive. The criteria which assure that P_{n-1} be a Hurwitz polynomial may, according to Eqs. 665 and 666, be expressed by writing

$$D_{n-1} = \begin{vmatrix} a_2 & a_1 & 0 & 0 & \cdots \\ a_4 & a_3 & a_2 & a_1 & \cdots \\ a_6 & a_5 & a_4 & a_3 & \cdots \\ a_8 & a_7 & a_6 & a_5 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{vmatrix} > 0 \qquad [671]$$

with the understanding that $a_1 > 0$ and that one is to consider all principal minors of this determinant which are formed by the first, the first two, the first three, etc., rows and columns.

According to determinant theory it is readily recognized that

$$a_1 D_{n-1} = \begin{vmatrix} a_1 & 0 & 0 & 0 & 0 & \cdots \\ a_3 & a_2 & a_1 & 0 & 0 & \cdots \\ a_5 & a_4 & a_3 & a_2 & a_1 & \cdots \\ a_7 & a_6 & a_5 & a_4 & a_3 & \cdots \\ a_9 & a_8 & a_7 & a_6 & a_5 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{vmatrix} \qquad [672]$$

which is formed from the determinant 671 by the addition of a first row and column as indicated. Since $a_1 > 0$, the conditions formed from this modified determinant are identical with those formed from 671 except that one must begin with the consideration of two rows and columns in order to obtain successively the same conditions as before.

The first, third, fifth, etc., columns in the determinant 672 are now multiplied by the positive constant a_0, and the determinant is then modified in form (although not in value) by adding to the elements of the second column the corresponding ones of the first column, and to the elements of the fourth column the corresponding ones of the third column, and so forth. This yields

$$\begin{vmatrix} a_0a_1 & a_0a_1 & 0 & 0 & \cdots \\ a_0a_3 & (a_0a_3 + a_2) & a_0a_1 & a_0a_1 & \cdots \\ a_0a_5 & (a_0a_5 + a_4) & a_0a_3 & (a_0a_3 + a_2) & \cdots \\ a_0a_7 & (a_0a_7 + a_6) & a_0a_5 & (a_0a_5 + a_4) & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{vmatrix} \qquad [673]$$

If the constant a_0 is now imagined to be factored out of the first, third, etc., columns, one observes that the result is a determinant D_n from which the criteria for the polynomial 670 are formed in the same manner that the criteria for the polynomial 666 are formed from the determinant 671. Thus, by the method of induction, the correctness of the relations 664 and 665 for the desired criteria (these are the Hurwitz criteria) is established.

Routh, who first derived these criteria* (although he stated them in a somewhat modified form) made use of a theorem of Cauchy's, which, for the sake of its collateral interest, is briefly discussed in the following paragraphs.

A finite polynomial is written in the form

$$P(z) = u(x,y) + jv(x,y) \qquad [674]$$

**Adams Prize Essay*, 1877; *Rigid Dynamics*, paragraph 290.

and the behavior of the quotient

$$q(x,y) = \frac{u(x,y)}{v(x,y)} \tag{675}$$

is observed as the point $z = x + jy$ traverses once around a simple closed contour C which does not pass through a zero of $P(z)$. As is shown in detail presently, the function $q(x,y)$ is seen to change its algebraic sign either by passing continuously through zero or by passing through the value ∞. Considering only the continuous passages through zero, one observes the number of changes of sign from plus to minus and the number of changes from minus to plus. Denoting these numbers by r and s respectively, the theorem of Cauchy states that $\frac{1}{2}(r - s)$ is the number of zeros of $P(z)$ enclosed by C.

To prove this theorem it is expedient to write the polynomial in the factored form

$$P(z) = A(z - z_1)(z - z_2) \cdots (z - z_n) \tag{676}$$

in which A is a real constant and is assumed to be positive. Each factor may be represented in the polar form

$$(z - z_\nu) = r_\nu e^{j\theta_\nu} \tag{677}$$

whence

$$P(z) = A \cdot r_1 \cdot r_2 \cdots r_n \cdot e^{j(\theta_1 + \theta_2 + \cdots + \theta_n)} \tag{678}$$

Letting

$$A \cdot r_1 \cdot r_2 \cdots r_n = R \tag{679}$$

and

$$\theta_1 + \theta_2 + \cdots + \theta_n = \theta \tag{680}$$

one has

$$P(z) = Re^{j\theta} = R\cos\theta + jR\sin\theta \tag{681}$$

Reference to Eqs. 674 and 675 then shows that

$$q(x,y) = \cot\theta \tag{682}$$

It remains to see how the angle θ behaves as the variable $z = x + jy$ traverses once around a simple closed contour. For a single factor, as given by Eq. 677, this behavior is illustrated in Fig. 50. In part (a) of this figure the point $z = z_\nu$ is assumed to lie outside the closed contour C, whereas in part (b) of the figure the point $z = z_\nu$ lies within the contour. It is immediately apparent that the *net* change in θ_ν as z traverses the contour is either zero or 2π according to whether z_ν lies outside or inside the contour. The net change in the resultant angle θ, as given by Eq. 680, is

thus seen to be $2\pi N$, with N equal to the number of zeros of $P(z)$ enclosed by the contour C. Reference to the expression 682 for the quotient u/v now reveals the truth of the theorem without further difficulty. [694]

It may be of interest to point out that the behavior of the angle θ can be determined alternatively from the theorem given in Art. 19 involving

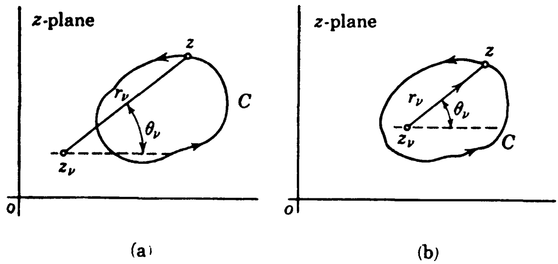


FIG. 50. The *net* change of θ_ν is 0 for (*a*) but is 2π for (*b*)

the number of zeros and poles of a function within a given region. Let the function be $f(z)$, and write it in the polar form

$$f(z) = |f| e^{j\theta} \qquad [683]$$

Then

$$\theta = Im\ (\ln f) \qquad [684]$$

in which Im denotes "imaginary part of." Now

$$d\theta = Im\left(\frac{df}{f}\right) \qquad [685]$$

and hence Eq. 282 yields

$$\oint_C d\theta = \Delta\theta = Im \oint_C \frac{f'(z)}{f(z)}\, dz = 2\pi(N - P) \qquad [686]$$

Thus the net change in angle $\Delta\theta$ is found to be equal to 2π times the difference between the number of zeros and the number of poles of $f(z)$ enclosed by the contour C. The polynomial $P(z)$, given by Eqs. 674 or 676, can have no poles in the finite z-plane, and hence the present result agrees with the conclusion reached above.

27. Positive real functions

A rational function $w = f(z)$ which is real for real values of z, and whose real part is positive for all values of z with a positive real part, is called a

and the [illegible] function* (abbreviated p.r. function). These conditions are

$$\begin{array}{ll} f(z) \text{ real} & \text{for } z \text{ real} \\ Re[f(z)] \geqq 0 & \text{for } Re(z) \geqq 0 \end{array} \qquad [687]$$

Functions of this sort play an important part in electrical network theory, and it is therefore of interest to study their properties in some detail.

Since the function very often is an impedance, it is customarily denoted by the letter Z. The independent variable is the complex frequency, which is written $\lambda = \sigma + j\omega$. Being a rational function, $Z(\lambda)$ is given by the quotient of two polynomials, and since $Z(\lambda)$ is real for real values of λ, these polynomials must have real coefficients. One may write

$$Z(\lambda) = \frac{P(\lambda)}{Q(\lambda)} = \frac{u_1(\sigma,\omega) + jv_1(\sigma,\omega)}{u_2(\sigma,\omega) + jv_2(\sigma,\omega)} \qquad [688]$$

in which u_1 and u_2 are the real parts of the polynomials P and Q respectively, and v_1, v_2 are the corresponding imaginary parts. Rationalizing, one finds

$$Re[Z(\lambda)] = \frac{u_1u_2 + v_1v_2}{u_2^2 + v_2^2} \qquad [689]$$

and similarly

$$Re\left[\frac{1}{Z(\lambda)}\right] = \frac{u_1u_2 + v_1v_2}{u_1^2 + v_1^2} \qquad [690]$$

Since the denominators in the last two expressions must always be positive, it is evident that if $Z(\lambda)$ is positive real, *its reciprocal is also positive real.*

The poles of $Z(\lambda)$ are the zeros of $Q(\lambda)$. In the [illegible] *inity of one of these, one may represent $Z(\lambda)$ by the Laurent series*

$$[illegible] \qquad [691]$$

[illegible] write

$$[illegible] \qquad [692]$$

and

$$b_{-s} = ke^{j\beta} \qquad [694]$$

one obtains for this vicinity

$$Re[Z(\lambda)] \cong \frac{k}{\rho^s} \cos (s\phi - \beta) \qquad [695]$$

As the immediate neighborhood of the pole is explored through allowing ϕ to vary from zero to 2π, one observes that the real part of $Z(\lambda)$ changes sign $2s$ times. It is clear, therefore, that $Z(\lambda)$ can have no poles in the right half of the λ-plane. It follows furthermore that poles may lie on the imaginary axis but that such poles must be simple ($s = 1$) and the function $Z(\lambda)$ must there have positive real residues ($\beta = 0$) so that, for the immediate vicinity of such a pole, Eq. 695 becomes

$$Re[Z(\lambda)] \cong \frac{k}{\rho} \cos \phi \qquad [696]$$

which remains positive for $-\pi/2 < \phi < \pi/2$, that is, for values of λ in the right half plane.

Since the same conclusions apply also to the reciprocal function $1/Z(\lambda)$, one recognizes that the zeros of the polynomials $P(\lambda)$ and $Q(\lambda)$ must have real parts which are not positive. *That is, $P(\lambda)$ and $Q(\lambda)$ must be Hurwitz polynomials.**

More explicitly, the expression for $Z(\lambda)$ may be written

$$Z(\lambda) = \frac{P(\lambda)}{Q(\lambda)} = \frac{a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_n\lambda^n}{b_0 + b_1\lambda + b_2\lambda^2 + \cdots + b_m\lambda^m} \qquad [697]$$

If the degree n of $P(\lambda)$ is higher than the degree m of $Q(\lambda)$, then $Z(\lambda)$ has a pole at $\lambda = \infty$. Since this point may be regarded as lying on the *imaginary axis, such a pole, if present, must be simple. Similarly, if* $a_0 \neq 0$, the functi[illegible] at $\lambda = 0$ if $b_0 = 0$; it has a *pole of second order at* [illegible] *forth. Since the* point $\lambda = 0$ lies on t[illegible] imag[illegible] [illegible]lso be simple. *Recognizing that the same c*[illegible] [illegible]ocal of $Z(\lambda)$, one sees that the po[illegible] [illegible] further restriction that *the lo*[illegible] [illegible]*nomials* $P(\lambda)$ *and* $Q(\lambda)$ *can diff*[illegible]

In examining a give[illegible] [illegible]ing whether or not it is p[illegible] [illegible] that $P(\lambda)$ and $Q(\lambda)$ are Hu[illegible]itz pol[illegible] [illegible] well as their highest powers differ by n[illegible] [illegible]pres-[illegible] [illegible]having

*The term *Hurwitz polynomial* as used i[illegible] zeros on the imaginary axis.

sion 689 for the real part of $Z(\lambda)$ and examine its behavior over the entire right half of the λ-plane, however, is a laborious procedure which one would like to avoid. In this regard it is found helpful to make use of the theorem (discussed in Art. 21) which states that if a function is analytic within and on the boundary of a given region, the maximum and minimum values which the real and imaginary parts of that function assume on the boundary are maxima and minima for the enclosed region. This region, in the present problem, is taken to be the right half of the λ-plane; its boundary is the imaginary axis.

If $Z(\lambda)$ is a positive real function, it is analytic in the right half plane, and if, for the moment, one assumes that $Z(\lambda)$ has no poles on the imaginary axis, the theorem just cited assures that the smallest value which the real part of $Z(\lambda)$ assumes on the imaginary axis must be smaller than any value which this real part may have over the entire right half plane. Conversely, if $Z(\lambda)$ is analytic in the right half plane and on the imaginary axis, and if the real part of $Z(\lambda)$ on the imaginary axis is nowhere negative, this real part must remain positive over the entire right half plane, and $Z(\lambda)$ must be a positive real function.

The stipulation that $Z(\lambda)$ be analytic on the imaginary axis may be dispensed with, for if $Z(\lambda)$ has poles on the imaginary axis, it is merely necessary to modify this boundary by inserting vanishingly small semicircular detours at such poles, so that the resulting boundary avoids these points by passing slightly to the right of them. As shown above, the requirement that the real part of $Z(\lambda)$ shall remain positive on a small semicircular detour is taken care of by the stipulation that poles on the imaginary axis be simple and that the residues of $Z(\lambda)$ at such poles may be real and positive.

The necessary and sufficient conditions that a rational function $Z(\lambda)$, which is real for real values of λ, be positive real may thus be stated in a form which does not require an investigation of the real part of $Z(\lambda)$ over the entire right half plane. Such a statement reads:

> *If* $Z(\lambda)$ *is analytic in the right half plane, and if on the imaginary axis this function has only simple poles with positive real residues,* $Z(\lambda)$ *is a positive real function if* $\mathrm{Re}[Z(j\omega)] \geqq 0$ *for all real values of* ω. [698]

In order to form the real part of $Z(\lambda)$ for $\lambda = j\omega$, one may begin with the expression 697 and in each of the two polynomials group the terms with even and odd powers respectively. That is, the polynomials are written

$$\begin{aligned} P(\lambda) &= m_1(\lambda) + n_1(\lambda) \\ Q(\lambda) &= m_2(\lambda) + n_2(\lambda) \end{aligned} \qquad [699]$$

in which m_1 and m_2 are the terms involving even powers of λ, whereas n_1 and n_2 are the terms involving odd powers of λ. It is clear that for $\lambda = j\omega$, m_1 and m_2 are real, whereas n_1 and n_2 are imaginary. A process of rationalization applied to $Z(\lambda)$, therefore, yields

$$Z(\lambda) = \frac{(m_1 + n_1)(m_2 - n_2)}{(m_2 + n_2)(m_2 - n_2)} \tag{700}$$

from which it is clear that*

$$Re[Z(j\omega)] = \left(\frac{m_1 m_2 - n_1 n_2}{m_2^2 - n_2^2}\right)_{\lambda = j\omega} \tag{701}$$

The denominator in this expression represents the square of an absolute value and hence is surely positive. The condition $Re[Z(j\omega)] \geqq 0$ is thus seen to be expressed by

$$(m_1 m_2 - n_1 n_2) \geqq 0, \quad \text{for } \lambda = j\omega \tag{702}$$

If $P(\lambda)$ and $Q(\lambda)$ are assumed to have the same degree n (this assumption does not restrict the generality of the present argument), it is clear that $(m_1 m_2 - n_1 n_2)$ is a polynomial of the degree n in the variable λ^2. One may, therefore, write

$$(m_1 m_2 - n_1 n_2) = A(\lambda_1^2 - \lambda^2)(\lambda_2^2 - \lambda^2) \cdots (\lambda_n^2 - \lambda^2) \tag{703}$$

or, for $\lambda = j\omega$,

$$(m_1 m_2 - n_1 n_2) = A(\lambda_1^2 + \omega^2)(\lambda_2^2 + \omega^2) \cdots (\lambda_n^2 + \omega^2) \tag{704}$$

The constant A must evidently be positive if this expression is to be positive for all values of ω, since it must still be positive for $\omega \to \infty$. The λ^2-roots, which are denoted by $\lambda_1^2, \lambda_2^2, \cdots \lambda_n^2$, may be complex as well as real, but since the polynomial 703 has real coefficients, any complex roots, if present, must occur in conjugate pairs. Such a pair of roots leads to a pair of conjugate complex factors in the expression 704, and hence yields a resultant factor which is the square of an absolute value. Complex as well as positive real λ^2-roots, therefore, yield factors in the expression 704 which are surely positive for all real values of ω. This statement is still true if some of the real λ^2-roots are zero. A negative real λ^2-root of even multiplicity leads to a factor in the expression 704 which is raised to an even power, and hence such a factor is also surely positive. However, if there exists a negative real λ^2-root of odd multiplicity (for example, a simple root of this sort), the expression 704 is surely negative over some part of the range $0 < \omega^2 < \infty$.

*The functions $m(\lambda)$ and $n(\lambda)$ should not be confused with $u(\sigma,\omega)$ and $v(\sigma,\omega)$ appearing in Eqs. 688, 689, and 690. m and n become identical with u and jv respectively for $\lambda = j\omega$, that is, only for $\sigma = 0$.

It thus becomes clear that *the necessary and sufficient condition insuring* $\mathrm{Re}[Z(j\omega)] \geqq 0$ *is simply that, in addition to being positive for* $\lambda^2 = -\infty$, *the polynomial* $(m_1m_2 - n_1n_2)$ *shall have no negative real* λ^2*-roots of odd multiplicity*. The fulfillment of this condition together with an assurance that $Q(\lambda)$ is a Hurwitz polynomial suffices to prove that $Z(\lambda)$ is a positive real function provided $Q(\lambda)$ has no zeros on the imaginary axis. If it has, one must also establish that these zeros are simple and that the residues of $Z(\lambda)$ there are real and positive. If $Z(\lambda)$ is found to be a positive real function, one may be sure that $P(\lambda)$ is, of course, also a Hurwitz polynomial.

If $Q(\lambda)$ has zeros on the imaginary axis and $P(\lambda)$ does not, it is easier to establish the positive real character of the reciprocal function $1/Z(\lambda)$ because the latter then has no poles on the imaginary axis although $Z(\lambda)$ does. When $1/Z(\lambda)$ is found to be positive real, it follows without further proof that $Z(\lambda)$ is positive real also.

A limiting form of the positive real function $Z(\lambda)$ results if its real part is identically zero for $\lambda = j\omega$. According to Eq. 701 this situation requires the condition

$$(m_1m_2 - n_1n_2) \equiv 0 \qquad [705]$$

in which λ need not be restricted to pure imaginary values.* If, for the moment, none of the polynomials m_1, m_2, n_1, n_2 are considered to be zero, one may write the condition 705 in the form

$$n_1 \equiv \frac{m_1m_2}{n_2} \qquad [706]$$

a representation for the polynomial n_1 that is possible only if m_1m_2 contains n_2 as a factor. Since $m_2 + n_2$ is a Hurwitz polynomial, the parts m_2 and n_2 have no common factors (the discussion of the preceding article shows that the zeros of m_2 and n_2 separate each other on the imaginary axis). Hence the identity 706 can be fulfilled only if

$$m_1 \equiv n_2\lambda^p \qquad \text{and} \qquad n_1 \equiv m_2\lambda^p \qquad [707]$$

the factor λ^p arising from the observation that one may multiply numerator and denominator on the right-hand side of the identity 706 by any power of λ. The functions m and n being respectively even and odd, it is clear that the integer p is odd.

The results 707 now yield for the impedance

$$Z(\lambda) = \frac{(m_2 + n_2)\lambda^p}{m_2 + n_2} = \lambda^p \qquad [708]$$

*It should not be inferred from Eq. 701 that, if $m_1m_2 - n_1n_2$ vanishes for all values of λ, so does the real part of $Z(\lambda)$, for the expression 701 represents the real part of $Z(\lambda)$ only for $\lambda = j\omega$, not for λ-values in the complex plane.

in which one can insert a constant multiplier if desired. Since a pole of $Z(\lambda)$ at infinity must be simple, one has $p = 1$. The function $Z(\lambda)$ thus obtained is trivially simple.

The only other conclusions permitted by the identity 705, if $Z(\lambda)$ is not to become identically zero or identically infinite, read

$$m_1 \equiv 0 \qquad \text{and} \qquad n_2 \equiv 0 \tag{709}$$

or

$$n_1 \equiv 0 \qquad \text{and} \qquad m_2 \equiv 0 \tag{710}$$

Correspondingly the impedance becomes

$$Z(\lambda) = \frac{n_1(\lambda)}{m_2(\lambda)} \tag{711}$$

or

$$Z(\lambda) = \frac{m_1(\lambda)}{n_2(\lambda)} \tag{712}$$

According to the discussion of Hurwitz polynomials given in the preceding article it is seen that these functions $Z(\lambda)$ have simple zeros and poles restricted to the imaginary axis. Thus the special form of a positive real $Z(\lambda)$-function whose real part is identically zero for $\lambda = j\omega$ is the same as that special function whose poles are restricted to the imaginary axis. One may say that if $Z(\lambda)$ is a positive real function whose real part is identically zero for $\lambda = j\omega$, the zeros of *both* $P(\lambda)$ and $Q(\lambda)$ must lie on the imaginary axis.

Since $Z(\lambda)$ is by definition a positive real function, the residues in all its poles are real and positive. A partial fraction expansion of such a $Z(\lambda)$-function reads

$$Z(\lambda) = \frac{k_0}{\lambda} + \frac{k_2}{\lambda - \lambda_2} + \frac{k_2}{\lambda + \lambda_2} + \frac{k_4}{\lambda - \lambda_4} + \frac{k_4}{\lambda + \lambda_4} + \cdots + k_{2p}\lambda \tag{713}$$

The terms in this expression, except for the first and last, are pairs of conjugates. Since the poles are restricted to the imaginary axis, the conjugate of λ_ν is $-\lambda_\nu$; and since the residues are real, conjugate poles (which normally involve conjugate residues because $Z(\lambda)$ must be real for real λ's) yield identical residues. The first and last terms in Eq. 713 represent possible poles at $\lambda = 0$ and $\lambda = \infty$.

Combining conjugate complex terms in the expression 713 gives

$$Z(\lambda) = \frac{k_0}{\lambda} + \frac{2k_2\lambda}{\lambda^2 - \lambda_2^2} + \frac{2k_4\lambda}{\lambda^2 - \lambda_4^2} + \cdots + k_{2p}\lambda \tag{714}$$

Since

$$\lambda_\nu = \pm j\omega_\nu \qquad \text{and} \qquad \lambda_\nu^2 = -\omega_\nu^2 \tag{715}$$

one finds

$$Z(j\omega) = \frac{k_0}{j\omega} + \frac{2k_2 j\omega}{\omega_2^2 - \omega^2} + \frac{2k_4 j\omega}{\omega_4^2 - \omega^2} + \cdots + k_{2p} j\omega \qquad [716]$$

This special form of the function $Z(j\omega)$ is of practical interest because it represents the impedance of a lossless network. The function $Z(\lambda)$ in this case is evidently on the borderline with respect to the property of being positive real.

In terms of the factored forms of the polynomials $P(\lambda)$ and $Q(\lambda)$, this function $Z(j\omega)$ has a representation of the form

$$Z(j\omega) = \frac{H(\omega^2{}_1 - \omega^2)(\omega^2{}_3 - \omega^2) \cdots (\omega^2{}_{2p-1} - \omega^2)}{j\omega(\omega^2{}_2 - \omega^2)(\omega^2{}_4 - \omega^2) \cdots (\omega^2{}_{2p-2} - \omega^2)} \qquad [717]$$

in which H is a positive real constant.

Even in this special limiting case the real part of $Z(\lambda)$ is zero only on the imaginary axis of the λ-plane. For complex λ-values one has

$$Z(\lambda) = R(\sigma,\omega) + jX(\sigma,\omega) \qquad [718]$$

in which R and X are the real and imaginary parts. According to the Cauchy-Riemann Eqs. 12 and 13,

$$\frac{\partial R}{\partial \sigma} = \frac{\partial X}{\partial \omega} \qquad \frac{\partial R}{\partial \omega} = -\frac{\partial X}{\partial \sigma} \qquad [719]$$

Since $R(\sigma,\omega)$ is identically zero for $\sigma = 0$ and positive for $\sigma > 0$, it follows that $\partial R/\partial\sigma$ is positive and hence that $\partial X/\partial\omega$ is positive for $\sigma = 0$. In other words, when all the poles of $Z(\lambda)$ lie on the imaginary axis and its real part there is identically zero, one has

$$\left(\frac{dZ}{d\lambda}\right)_{\lambda = j\omega} = \frac{dZ(j\omega)}{j\,d\omega} > 0 \quad \text{for } -\infty < \omega < \infty \qquad [720]$$

A similar relation then also holds for the reciprocal function.

This result states that the real function $Z(j\omega)/j$ of the real variable ω has a positive slope for all values of ω. As a consequence it follows that the zeros and poles of $Z(\lambda)$, which are all simple and lie along the imaginary axis, must mutually separate each other. This alternation of zeros and poles may (with reference to the notation in Eq. 717) be expressed by

$$0 < \omega_1 < \omega_2 < \cdots < \omega_{2p-2} < \omega_{2p-1} < \infty \qquad [721]$$

It is significant to observe that a positive real function in this limiting case has properties identical with those of the function $\psi(z)$ discussed in the previous article (see Eq. 625 and following). A function of this type is represented by the reactance or susceptance of a lossless electrical network. One may, therefore, state that the polynomials whose ratio

represents the reactance or the susceptance of a lossless network are the even and odd parts of a Hurwitz polynomial. The converse of this statement is likewise true. When the expression 688 represents such a reactance or susceptance function, one of the polynomials $P(\lambda)$ or $Q(\lambda)$ is an even and the other an odd function of λ, as shown by Eqs. 711 and 712.

Returning to the general case, one may discover additional useful properties of the positive real function by considering its linear fractional transformation

$$z(\lambda) = \frac{1 - Z(\lambda)}{1 + Z(\lambda)} = \frac{Q(\lambda) - P(\lambda)}{Q(\lambda) + P(\lambda)} \tag{722}$$

As shown in Art. 24, this transformation maps the right half of the Z-plane upon the interior of the unit circle in the z-plane (and vice versa), the imaginary axis of the Z-plane becoming identified with the unit circle in the z-plane. If $Z(\lambda)$ is a positive real function, points in the right half of the λ-plane correspond to points in the right half of the Z-plane and hence to points within the unit circle of the z-plane. One may state

$$\textit{If } Z(\lambda) \textit{ is positive real, } |z(\lambda)| \leqq 1 \textit{ for } \mathrm{Re}(\lambda) \geqq 0 \tag{723}$$

It is readily appreciated that the converse must be true also; that is,

$$\textit{If } |z(\lambda)| \leqq 1 \textit{ for } \mathrm{Re}(\lambda) \geqq 0, \textit{ then } Z(\lambda) \textit{ is positive real} \tag{724}$$

Moreover, if $Z(\lambda)$ is positive real, then $z(\lambda)$ must be analytic in the right half plane and on the imaginary axis, for $1 + Z(\lambda)$ cannot be zero there. According to the principle of the maximum modulus (discussed in Art. 21), the largest value which $|z(\lambda)|$ assumes on the imaginary axis of the λ-plane must be a maximum for the entire right half of that plane.

Conversely, one may state that if $z(\lambda)$ in Eq. 722 can be shown to be analytic in the right half plane inclusive of the imaginary axis, and if on this boundary $|z(\lambda)| \leqq 1$, this inequality must hold for the entire right half plane and, according to the statement 724, $Z(\lambda)$ must be positive real.

It is now interesting to consider the function

$$Z^*(\lambda) = \frac{P^*(\lambda)}{Q^*(\lambda)} \tag{725}$$

in which

$$\begin{aligned} P^*(\lambda) &= m_1(\lambda) + n_2(\lambda) \\ Q^*(\lambda) &= m_2(\lambda) + n_1(\lambda) \end{aligned} \tag{726}$$

Comparing these relations with Eqs. 699, one observes that the polynomials $P^*(\lambda)$ and $Q^*(\lambda)$ are formed from $P(\lambda)$ and $Q(\lambda)$ by interchange of their odd parts. In terms of $Z^*(\lambda)$ one now forms the expression

analogous to 722, namely,

$$z^*(\lambda) = \frac{1 - Z^*(\lambda)}{1 + Z^*(\lambda)} = \frac{Q^*(\lambda) - P^*(\lambda)}{Q^*(\lambda) + P^*(\lambda)} \tag{727}$$

According to Eqs. 699 and 726 it is readily seen that

$$Q^*(\lambda) + P^*(\lambda) = Q(\lambda) + P(\lambda) \tag{728}$$

Since it has already been shown that the polynomial $P(\lambda) + Q(\lambda)$ has no zeros in the right half plane or on the imaginary axis when $Z(\lambda)$ is positive real, one observes in that event that $z^*(\lambda)$ is analytic in the entire right half plane inclusive of the imaginary axis.

From the mapping properties of the function 727 it follows that $|z^*(j\omega)| \leqq 1$ for $Re[Z^*(j\omega)] \geqq 0$. The latter inequality is found, from Eqs. 725 and 726, to be expressed by

$$\left(\frac{m_1m_2 - n_1n_2}{m_2^2 - n_1^2}\right)_{\lambda = j\omega} \geqq 0 \tag{729}$$

or, since $(m_2^2 - n_1^2)$ for $\lambda = j\omega$ is the square of an absolute value, one has more simply

$$(m_1m_2 - n_1n_2) \geqq 0, \quad \text{for } \lambda = j\omega \tag{730}$$

This is the condition 702, which is fulfilled if $Z(\lambda)$ is positive real. In this event the present argument is thus seen to yield $|z^*(j\omega)| \leqq 1$, and since $z^*(\lambda)$ is analytic in the right half plane inclusive of the imaginary axis, the principle of the maximum modulus enables one to conclude that $|z^*(\lambda)| \leqq 1$ throughout the entire right half plane. According to the statement 724, therefore, the function $Z^*(\lambda)$ is then seen to be positive real. This conclusion is summarized by the statement:

If the rational function $Z(\lambda)$, *given by the quotient of polynomials* $P(\lambda)$ *and* $Q(\lambda)$, *is positive real, the rational function which results after the even or odd parts of these polynomials are interchanged is also positive real.* [731]

Since $Z^*(\lambda)$ is positive real, it follows that $P^*(\lambda)$ and $Q^*(\lambda)$ are Hurwitz polynomials. Hence one may state

If the quotient of polynomials $P(\lambda)/Q(\lambda)$ *is a positive real function, not only these polynomials but also those which result from an interchange of their even or odd parts are Hurwitz polynomials.* [732]

Further practically useful results are obtained from an investigation of the properties of positive real functions in polar form. The first step in this direction is the introduction of the relations

$$z(\lambda) = \frac{\lambda - A}{\lambda + A} \quad \text{or} \quad \lambda = A \cdot \frac{1 + z}{1 - z} \tag{733}$$

and

$$w(z) = \frac{Z(\lambda) - B}{Z(\lambda) + B} \tag{734}$$

The positive real constants A and B have any values satisfying the relation

$$Z(A) = B \tag{735}$$

By means of the transformation 733 the interior of the unit circle in the z-plane is mapped upon the right half of the λ-plane. The positive real function $Z(\lambda)$ relates points in the right half λ-plane to points in the right half Z-plane, and these in turn are mapped upon the interior of the unit circle in the w-plane by means of the transformation 734. As a consequence of the relation 735 between A and B, the origin in the z-plane corresponds to the origin in the w-plane, that is,

$$w(0) = 0 \tag{736}$$

The function Z and its independent variable λ are now represented in the polar forms

$$\lambda = \rho e^{j\phi} \quad \text{and} \quad Z = re^{j\theta} \tag{737}$$

In terms of the variables thus introduced, Fig. 51 illustrates the identical transformations 733 and 734. Part (a) shows how the concentric circles (ρ = constant) and the radial lines (ϕ = constant) of the right half λ-plane appear inside the unit circle of the z-plane, and part (b) similarly illustrates the appearance of the polar representation of Z within the unit circle of the w-plane. In each case the unit circle itself represents the imaginary axis of the λ- or Z-plane, the left-hand point on the circle ($z = w = -1$) corresponding to the origin in either of these planes and the right-hand one ($z = w = 1$) to the point at infinity.

The imaginary axis of the λ-plane thus corresponds to $|z| = 1$. Therefore, if $Z(\lambda)$ is a positive real function, points on the locus $|z| = 1$ yield Z-values which are in the right half or on the imaginary axis of the Z-plane and hence within or on the unit circle of the w-plane. That is, one may state that

$$|w| \leqq 1 \quad \text{for } |z| = 1 \tag{738}$$

In view of this result and the condition $w(0) = 0$, as expressed by

Eq. 736, it is recognized that Schwarz's lemma (see Art. 21) enables one to make a considerably stronger statement, namely, that

$$|w| \leqq |z| \quad \text{for } |z| < 1 \tag{739}$$

in which the equals sign holds only if it holds identically.

This result is readily translated into an expression involving λ and $Z(\lambda)$ since it states, with reference to Fig. 51, that to any concentric circle within the unit circle of the z-plane there corresponds a concentric

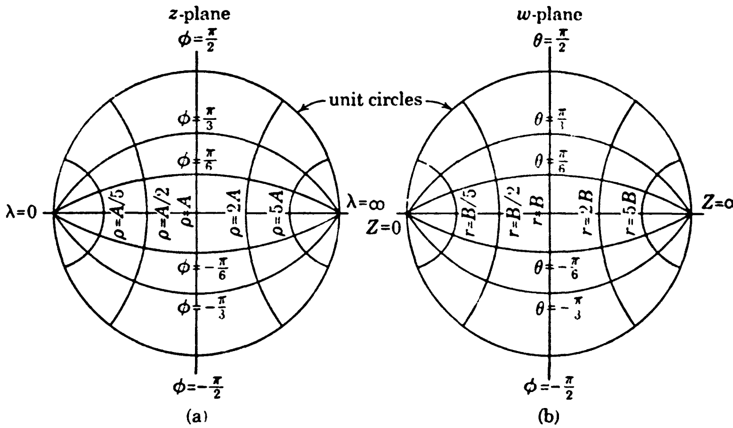


FIG. 51. Relevant to the derivation of the properties of a p. r. function expressed in polar form.

circle within the unit circle of the w-plane which is at least as small or smaller. If one considers the value $\rho = A$, which implies no restriction since the value of A is arbitrary, one observes by inspection of the figures that the following condition obtains:

$$|\theta| \leqq |\phi| \quad \text{for } |\phi| \leqq \frac{\pi}{2} \tag{740}$$

This result may be written in the alternate form

$$|\arg. Z| \leqq |\arg. \lambda| \quad \text{for } 0 < |\arg. \lambda| \leqq \frac{\pi}{2} \tag{741}$$

Again the equals sign in the first of these inequalities holds only if it holds identically.

The remarkable part about this result is that although it appears to be stronger, it nevertheless is contained in the statement $Re[Z(\lambda)] \geqq 0$ for $Re(\lambda) \geqq 0$, expressing the positive real character of $Z(\lambda)$. Since the present result is readily seen to include this statement in terms of the

real parts of λ and $Z(\lambda)$, one concludes that the two forms of expressing the positive real character of $Z(\lambda)$ are entirely equivalent. The statement in terms of the angles of λ and $Z(\lambda)$, just derived, is a translation of the one in terms of real parts into its equivalent polar form.

An additional inequality may be obtained from these considerations. With reference to the transformation 733 and the polar representation 737 for λ, let $\rho = A$. This choice is always possible since the value of A is arbitrary. Equation 733 then yields

$$z = \frac{e^{j\phi} - 1}{e^{j\phi} + 1} = j \tan \frac{\phi}{2} \qquad [742]$$

The condition 739 may therefore be written

$$|w| \leqq \tan \frac{\phi}{2} \quad \text{for } 0 < \phi < \frac{\pi}{2} \qquad [743]$$

Using the inverse of the transformation 734, which reads

$$\frac{Z}{B} = \frac{1 + w}{1 - w} \qquad [744]$$

and taking note of Eq. 735 ,one obtains the condition

$$\frac{1 - \tan \frac{\phi}{2}}{1 + \tan \frac{\phi}{2}} \leqq \frac{|Z(\lambda)|}{Z(\rho)} \leqq \frac{1 + \tan \frac{\phi}{2}}{1 - \tan \frac{\phi}{2}}, \quad \text{for } 0 < \phi < \frac{\pi}{2} \qquad [745]$$

For $\phi = \pi/2$ the lower and upper limits in this relation are zero and infinity, but for $\phi = \pi/4$, for example, one has

$$0.414 \leqq \frac{|Z(\lambda)|}{Z(\rho)} \leqq 2.414 \qquad [746]$$

The restriction implied by the condition 745 is thus seen to illustrate another interesting property of positive real functions.

A property which is rather obvious, but nevertheless practically useful, is expressed by the statement:

If Z(λ) *and* W(λ) *are positive real functions,* f(λ) = Z(W) *is again positive real.* [747]

In other words, a positive real function of a positive real function is also positive real. As an example one may consider the simple positive real function

$$W(\lambda) = \frac{1}{\lambda} \qquad [748]$$

The above statement asserts that if $Z(\lambda)$ is positive real, $Z(1/\lambda)$ is also positive real. Again, one may have

$$W(\lambda) = \lambda + \frac{1}{\lambda} \tag{749}$$

which is readily recognized as being positive real. Then if $Z(\lambda)$ is positive real, one may state at once that $Z(\lambda + [1/\lambda])$ is likewise positive real. Use of the condition 741 enables one to make the further assertion that the angle of the function $Z(\lambda + [1/\lambda])$ must, for every value of λ in the right half plane, be smaller than the angle of $Z(\lambda)$.

PROBLEMS

1. Verify that the real and imaginary parts of the function

$$e^x \cos y + je^x \sin y$$

satisfy the Cauchy-Riemann equations. Show that this function must be e^z where $z = x + jy$.

2. Using this reasoning, show that

$$\sin x \cosh y + j \cos x \sinh y = \sin z$$

$$\cos x \cosh y + j \sin x \sinh y = \cos z$$

3. Let the sphere in stereographic projection be of radius 1, so that its equation is

$$\xi^2 + \eta^2 + \zeta^2 - 2\zeta = 0$$

Show that the point ξ, η, ζ on the sphere corresponds to the point

$$x = \frac{2\xi}{2 - \zeta} \qquad y = \frac{2\eta}{2 - \zeta}$$

in the plane $\zeta = 0$. Show that if $z_2\bar{z}_1 = -4$, z_1 and z_2 correspond to diametrically opposite points.

4. For $w = \sin z = u + jv$, sketch the curves $u =$ constant and $v =$ constant in the z-plane and verify that they are orthogonal.

5. Actually integrate $z^2\,dz$ around the following contours, and verify that the results are zero: (a) Around the square having vertices at $1 + j$, $1 - j$, $-1 - j$, $-1 + j$. (b) Around the triangle having vertices at $1 + j$, $-1 - j$, $-1 + j$.

6. Carry out the integration for $(1/z)\,dz$ about the following contours: (a) The square with vertices at $1, 2, 2 + j, 1 + j$. (b) A circle of radius 1 about the origin as center. Why is the result not zero in part (b)?

7. As an example of analytic continuation, consider the following: (a) Find the function represented by $1 + z + z^2 \cdots$. (b) Determine its circle of convergence, and find the singularity on the circumference. (c) Determine the series for this function about the point $z = -\tfrac{1}{2}$. (d) Verify that this series converges for points for which the original series diverges, in particular at the point $z = -\tfrac{3}{2}$.

8. As an example of a function having a natural boundary, consider

$$f(z) = 1 + z^2 + z^4 + z^8 + z^{16} + \cdots$$

(a) Show that the radius of convergence is 1. (b) Verify that for every point of the form $e^{2\pi jk/2^p}$, k and p being integers, all terms after z^{2^p} will be 1, so that the function is singular at such points. (c) Show that these points are dense on the unit circle, in the sense that there is no interval free of them. As a consequence the function cannot be continued outside.

9. In deriving the Laurent series for a function, the integral formula for the coefficients is not practical. The function

$$\frac{\sin z}{z^k}$$

has a pole of order $k - 1$ at the origin, and its Laurent series would be found by dividing the series for $\sin z$ by z^k. Using this idea, find the Laurent series for

$$\frac{\cos \pi z}{(1 - z)^2}$$

about the point $z = 1$.

10. In deriving a Laurent series, the method of partial fractions is useful. For example find the Laurent series for

$$\frac{1}{z^2(1 - z)}$$

about $z = 0$ and $z = 1$. What are the regions of convergence?

11. Sketch the lines u = constant and v = constant near the origin for the functions

$$w = 1 + z^4 \qquad \text{and} \qquad w = 1 + z^5$$

which have saddle points there.

12. Find the value of the integral of $\csc z\, dz$ taken around the following contours: (a) A circle of radius 1 and center at the origin. (b) A circle of radius 4 and center at the origin. (c) A circle of radius 2 and center at $z = 2$.

13. Find the integral of $\sin z/z^k$ for $k = 1, 2, 3$, and 4 taken about a circle of radius 1 and center at the origin.

14. Determine the polar and rectangular forms of the function

$$f(z) = \frac{\sinh nz}{n \sinh z}$$

in which n is an integer, and show that the Cauchy-Riemann equations are satisfied.

15. Show that $f(z)$ in the foregoing problem is an entire function, and determine the distribution of its zeros.

16. Suppose the derivative of $f(z) = u(x,y) + jv(x,y)$ is written

$$\frac{df}{dz} = a + jb = \sqrt{a^2 + b^2}\, e^{j \tan^{-1} b/a}$$

If $f(z)$ satisfies the Cauchy-Riemann condition equations, show that one may have

$$a = \frac{\partial u}{\partial x} \quad \text{and} \quad b = \frac{\partial v}{\partial x} \qquad \text{or} \qquad a = \frac{\partial v}{\partial y} \quad \text{and} \quad b = -\frac{\partial u}{\partial y}$$

For the following functions

$$z^n; \qquad \sqrt[n]{z}; \qquad \cosh z; \qquad \sinh z; \qquad e^{jz}; \qquad \ln z; \qquad \cos^{-1} z$$

compute the polar form of df/dz through first forming u and v and obtaining a and b by one or the other of the above pairs of relations, and alternately through forming df/dz directly and then putting the result into its polar form. Check the three results for each function.

17. Derive the following identities:

(a) $\sinh jz = j \sin z$

(b) $\sin jz = j \sinh z$

(c) $\cosh jz = \cos z$

(d) $\cos jz = -\cosh z$

(e) $\tanh jz = j \tan z$

(f) $\tan jz = j \tanh z$

(g) $\sinh^{-1} jz = j \sin^{-1} z$

(h) $\sin^{-1} jz = j \sinh^{-1} z$

(i) $\cosh^{-1} z = j \cos^{-1} z$

(j) $\cosh^{-1} jz = j \cos^{-1} jz$

(k) $\tanh^{-1} jz = j \tan^{-1} z$

(l) $\tan^{-1} jz = j \tanh^{-1} z$

(m) $\sinh^{-1} z = \ln (z + \sqrt{z^2 + 1})$

(n) $\cosh^{-1} z = \ln (z + \sqrt{z^2 - 1})$

(o) $\tanh^{-1} z = \dfrac{1}{2} \ln \left(\dfrac{1 + z}{1 - z}\right)$

(p) $\coth^{-1} z = \dfrac{1}{2} \ln \left(\dfrac{z + 1}{z - 1}\right)$

(q) $\sin^{-1} z = \dfrac{1}{j} \ln (jz + \sqrt{1 - z^2})$

(r) $\cos^{-1} z = \dfrac{1}{j} \ln (z + \sqrt{z^2 - 1})$

(s) $\tan^{-1} z = \dfrac{1}{2j} \ln \left(\dfrac{1 + jz}{1 - jz}\right)$

(t) $\cot^{-1} z = \dfrac{1}{2j} \ln \left(\dfrac{jz - 1}{jz + 1}\right)$

18. Considering the function $w = \ln z$, draw in the w-plane the figure corresponding to the rectangle in the z-plane defined by

$$(x = -3;\ 1 \leqq y \leqq 2) \qquad (x = 3;\ 1 \leqq y \leqq 2)$$
$$(-3 \leqq x \leqq 3;\ y = 1) \qquad (-3 \leqq x \leqq 3;\ y = 2)$$

19. In a complex plane draw the figure bounded by arcs of concentric circles with radii $r_1 = 10$ and $r_2 = 30$ centimeters, and radial lines making angles of 30 and 60 degrees with respect to the positive real axis. Using the function $w = \ln z$, transform this figure into a rectangle and specify the equations of the straight lines forming its sides.

20. Considering the function $w = \sin^{-1} z = u + jv$, show that the lines $u =$ constant and $v =$ constant in the w-plane are transformed respectively into central hyperbolas

$$\frac{x^2}{\sin^2 u} - \frac{y^2}{\cos^2 u} = 1$$

and central ellipses

$$\frac{x^2}{\cosh^2 v} + \frac{y^2}{\sinh^2 v} = 1$$

in the z-plane, and demonstrate that these families are confocal and orthogonal. Draw a representative number of these loci.

21. With reference to the conformal map described in the previous problem, consider the strip $-\pi/2 \leq u \leq \pi/2$ in the w-plane and determine the regions in the z-plane corresponding to the portions of this strip defined by $v > 0$ and $v < 0$ respectively. The closed rectangular boundary in the w-plane joining the four points $u = 0$, $v = 0.002$; $u = 1$, $v = 0.002$; $u = 1$, $v = 0.02$; $u = 0$, $v = 0.02$ is traversed counterclockwise. Determine the corresponding contour in the z-plane and indicate the direction of traversal.

22. Continuing the study of the function $w = \sin^{-1} z$, describe the structure of its Riemann surface pointing out the location of branch points, appropriate positions of branch cuts, and the number of leaves in the surface. How many leaves of the Riemann surface are occupied by the closed contour in the z-plane corresponding to the rectangular one in the w-plane joining the four points:

$$u = -\frac{5\pi}{4}, v = 1; \quad u = \frac{5\pi}{4}, v = 1; \quad u = \frac{5\pi}{4}, v = 10; \quad u = -\frac{5\pi}{4}, \ v = 10$$

23. Consider the function $w = 1/z$ and show that it transforms circles in the z-plane into circles in the w-plane and vice versa (including straight lines as limiting forms of circles). Draw a family of circles in the w-plane corresponding to the straight lines $y = kx + c$ (with k and c real) in the z-plane, choosing a fixed value for k, for example, $k = \frac{1}{2}$, and various values for c as, for example, $c = 0, \pm 1, \pm 2$, etc. Alternately consider the families of straight lines $x =$ constant and $y =$ constant.

24. For the function $w = z/(a - z)$ with $a = 1 + j\frac{1}{2}$, map the loci in the z-plane corresponding to the lines $v = \frac{1}{2}u + 1, v = \frac{1}{2}u, v = \frac{1}{2}u - 1$ in the w-plane. Indicate corresponding directions of traversal by placing arrows on the loci. By shading indicate the regions in the z-plane corresponding to (a) that part of the w-plane below the line $v = \frac{1}{2}u$, (b) that part of the w-plane below $v = \frac{1}{2}u$ and above the real axis.

25. The arc of a circle of radius 10 and center in the first quarter of the z-plane passes through the points $x = 0$, $y = 0$ and $x = 3$, $y = -1$. Show that the region enclosed by this arc and the chord passing through the same two points may be mapped in the w-plane as an angular slit with its vertex at the origin by means of the transformation $w = z/(a - z)$. Determine the complex value of the constant a and calculate the angular aperture of the slit as well as its orientation in the w-plane.

26. A region A in the first quadrant of the z-plane is bounded by the arcs of three circles passing through the origin, two of them having their centers at the points $x = 1$, $y = 0$ and $x = 2$, $y = 0$, whereas the center of the third circle is at the point $x = 0$, $y = 2$. A second region B is bounded by the arcs of two circles passing through the origin with centers at $x = -2$, $y = 0$ and at $x = 0$, $y = -1$. Show that each region can be transformed into a rectangular strip in the w-plane through the transformation $w = 1/z$ and determine the boundaries of these corresponding rectangular regions.

27. Show that the equation

$$A(x^2 + y^2) + Bx + Cy + D = 0$$

in which the constants are subjected to the condition

$$B^2 + C^2 - 4AD > 0$$

is a general equation for circles and straight lines, and that this property is preserved when $z = x + jy$ is replaced by $1/z$.

28. For the transformation

$$w = \frac{1 - z}{1 + z}$$

let $z = re^{j\theta}$ and consider the concentric circles $r =$ constant and the radial lines $\theta =$ constant, and compute the location of the center and the radius of each corresponding circle in the w-plane in terms of r and θ respectively. Plot a representative family of curves in each plane.

29. For a function $w = f(z) = u + jv$, satisfying the Cauchy-Riemann equations, loci for $u =$ constant and $v =$ constant are plotted in the z-plane. If the slopes of these

curves are denoted by $(dy/dx)_{u=\text{constant}}$ and $(dy/dx)_{v=\text{constant}}$ respectively, show that at any point these derivatives have negative reciprocal values and hence that the families of curves intersect at right angles.

30. Study the function $w = e^{(z-a)(z-b)}$ in the vicinity of the point $z_0 = \frac{1}{2}(a + b)$ through plotting in this region a number of curves corresponding to u = constant and v = constant for $a = 5 + j0$ and $b = 0 + j5$.

31. Let $v_x = \partial u/\partial x$ and $v_y = \partial u/\partial y$ be the velocity components of a conservative hydrodynamic field, u being the real part of the analytic function $w = f(z)$. Show that the magnitude of the velocity is given by $v = |df/dz|$.

32. Write down the mutual conditions to be satisfied by two functions $P(x,y)$ and $Q(x,y)$ if the line integrals

$$\int_c (P\,dx - Q\,dy) \qquad \text{and} \qquad \int_c (Q\,dx + P\,dy)$$

are to be independent of the path c, and compare the results with the Cauchy-Riemann equations.

33. S is a path joining two points A and B in the complex z-plane, and $z_1, z_2, \cdots, z_n = B$ are n uniformly spaced points along this path. The definite integral along the path is defined as

$$\int_S f(z)\,dz = \underset{n\to\infty}{\text{limit}} \sum_{k=1}^{n} f(z_k)\,\Delta z_k, \quad \text{with } \Delta z_k = z_k - z_{k-1}$$

Show first that

$$\left|\sum_{k=1}^{n} f(z_k)\,\Delta z_k\right| \leqq \sum_{k=1}^{n} |f(z_k)| \cdot |\Delta z_k|$$

and thus that

$$\left|\int_S f(z)\,dz\right| \leqq \int_S |f(z)| \cdot |dz|$$

34. With reference to the situation given in Prob. 33, suppose $f(z)$ is bounded along S and that its largest absolute value on this path is M. Let the length of the path S between A and B be denoted by L. Show that

$$\left|\int_S f(z)\,dz\right| \leqq ML$$

35. Let $f(z)$ be regular on and within a closed contour C. The length of the contour is L, z is an internal point, and L' is the circumference of a circle whose radius is the minimum distance from z to the contour. By means of Cauchy's integral formula, show that

$$|f(z)| \leqq \frac{ML}{L'} \qquad (M \text{ defined in Prob. 34})$$

36. By direct integration along the sides of a rectangular contour joining the points (x,y); $(x + x_0,y)$; $(x + x_0,y + y_0)$; $(x,y + y_0)$ check the relation

$$\oint e^z\,dz = 0$$

37. Consider integration of the function e^z/z around a closed contour formed by two semicircles in the left half plane concentric at the origin, one having a large radius R, the other having a small radius ρ, and those portions of the imaginary axis joining the semicircles. The integral is effectively expressed as the sum of four parts corresponding to the two linear path increments and the two semicircular ones for which $z = Re^{j\theta}$ and $z = \rho e^{j\theta}$. Observing that the function e^z/z is regular on and within this contour, obtain in the limit $R \to \infty$, $\rho \to 0$, the result

$$\int_0^\infty \frac{\sin y}{y}\, dy = \frac{\pi}{2}$$

38. Compute the definite integral

$$\int_1^z \frac{d\zeta}{\zeta}$$

along a path consisting of the linear increment from the point 1 to the point $|z|$ on the positive real axis, followed by a circular increment concentric at the origin, expressing the result as the sum of two parts corresponding to these path increments. Thus show that the value of the integral equals $\ln z$.

39. Consider integration of the function $w = e^{-z^2}$ around a closed rectangular boundary joining the four points $z = -a$, $z = a$, $z = a + jb$, $z = -a + jb$, a and b having positive real values, and express this integral as the sum of four parts corresponding to the four linear path increments. Using the result

$$\int_0^\infty e^{-x^2}\, dx = \tfrac{1}{2}\sqrt{\pi}$$

obtain in the limit $a \to \infty$ the more general one

$$\int_0^\infty e^{-x^2} \cos 2bx\, dx = \frac{e^{-b^2}}{2}\sqrt{\pi}$$

40. Integrate the function e^{-z^2} around a closed contour formed by the linear increment from 0 to R along the positive real axis, followed by the circular arc from R to $Re^{j(\pi/4)}$, and completed through a linear increment from $Re^{j(\pi/4)}$ to 0. Considering the limit $R \to \infty$ and again using the value of the error integral given in Prob. 39, obtain the result

$$\int_0^\infty \cos x^2\, dx = \int_0^\infty \sin x^2\, dx = \frac{1}{2}\sqrt{\frac{\pi}{2}}$$

which, with the substitution $x^2 = u$, yields the Fresnel integrals.

41. Expand the following functions in Maclaurin series: e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$.

42. Through expansion in Maclaurin series, check the following expressions:

$$\ln(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots$$

$$(1 + z)^n = 1 + nz + \frac{n(n-1)}{1 \cdot 2} z^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} z^3 + \cdots$$

43. Using the result that the radius of the convergence circle of the power series

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n + \cdots$$

is given by

$$R = \lim_{n \to \infty} \left[\frac{1}{\sqrt[n]{|a_n|}} \right]$$

show that $R = \infty$ for the series in Prob. 41 and $R = 1$ for those in Prob. 42.

44. Find the radius of convergence of the following geometric series: $1 + z + z^2 + z^3 + \cdots + z^n + \cdots$.

45. Expand the function $(z^2 - 1)^{-2}$ in a power series about the points $z = 0$, $z = j$, $z = \frac{1}{2}$, and find the radius of convergence in each case.

46. Utilizing the fact that the convergence circle reaches to the nearest singularity, compute the radius of convergence of the function $[(z - 1)(z + 2 - j3)(z^2 + 9)]^{-1}$ expanded in a power series about each of the points: $z = 0$, $z = 0.2 - j3$, $z = j2$, $z = -8$.

47. Assuming that the function given in Prob. 46 is expanded in power series about points on the line $y = -2x$ lying at $x = 1, 0, -1, -2$, make sketches showing the various regions of convergence and their overlapping portions.

48. Show that the expansion of the function

$$f(z) = \int_0^\infty e^{-t(z-1)}\, dt$$

about appropriate points z_0 on the positive real axis reads

$$f(z) = \sum_{n=0}^{\infty} \frac{(z - z_0)^n(-1)^n}{(z_0 - 1)^{n+1}}$$

and verify that the radius of convergence is $R = |z_0 - 1|$.

49. Using the result

$$\int_0^\infty e^{-zt}\, dt = \frac{1}{z} \quad \text{for} \quad x > 0$$

find the values of the following integrals:

$$\int_0^\infty e^{-xt} \cos yt\, dt \qquad \int_0^\infty e^{-xt} \sin yt\, dt$$

50. Show that the function $1/(1 + \sqrt{z - 2})$ has two different power series expansions about the point $z = 0$ with radius of convergence equal to 1 and 2 respectively.

51. Using the formula for the coefficient b_n of the Laurent series, and assuming that the points ζ and z_0 are not functionally related, obtain the relation

$$b_{n+1} = \frac{-1}{n+1} \frac{\partial b_n}{\partial z_0}$$

applicable for negative as well as positive integers n.

52. Consider the rational function

$$R(z) = \frac{P_1(z)}{P_2(z)}$$

in which P_1 and P_2 are finite polynomials. Assume that $P_2(z)$ has a zero of multiplicity α at $z = z_0$ and write $P_2(z) = (z - z_0)^\alpha F_2(z)$. Letting $\zeta - z_0 = \rho e^{j\theta}$ in the formula

for Laurent coefficients and recognizing that ρ may become arbitrarily small, obtain the result

$$b_{-\alpha} = \frac{P_1(z_0)}{F_2(z_0)}$$

Also show that the formula yields $b_{-(\alpha+p)} = 0$ for all positive values of p, and using the result of Prob. 51, get

$$b_{-(\alpha-p)} = \frac{1}{(\alpha - 1)(\alpha - 2) \cdots (\alpha - p)} \frac{\partial^p}{\partial z_0{}^p} b_{-\alpha}$$

which contains the specific result

$$b_{-1} = \frac{1}{(\alpha - 1)!} \frac{\partial^{\alpha-1}}{\partial z_0^{\alpha-1}} b_{-\alpha}$$

53. Considering the Laurent expansion of the rational function described in Prob. 52, show that the coefficients $b_{-\alpha} \cdots b_{-1}$ have real values when z_0 is real and that they have conjugate complex values for conjugate values of z_0.

54. Show that the Laurent expansion of the function

$$f(z) = e^{(t/2)(z-[1/z])}$$

about its essential singularity at $z = 0$ has coefficients given by the formula

$$b_n(t) = \frac{1}{2\pi} \int_0^{2\pi} \cos (n\phi - t \sin \phi) \, d\phi$$

which are Bessel functions of the first kind.

55. Suppose $f(z)$ has an isolated singularity at $z = z_0$ and the radius of convergence for its Laurent expansion about this point is R. For points on a concentric circle about z_0 with radius $r < R$ show that the series representation takes the form

$$f(z) = A_0 + \sum_{n=1}^{\infty} \{(A_n + A_{-n}) \cos n\phi + j(A_n - A_{-n}) \sin n\phi\}$$

in which

$$A_n = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{j\phi})e^{-jn\phi} \, d\phi$$

56. Find Laurent expansions for the following rational functions about their poles:

$$\frac{5z^2 + 3z + 2}{4z^2 + 8z + 1} \qquad \frac{1}{4z^2 + 8z + 1} \qquad \frac{3z}{4z^2 + 8z + 1}$$

57. Find the partial fraction expansions of the functions given in Prob. 56.

58. A polynomial $P(z)$ has a double zero at $z = 1$, a zero at $z = j$, and a zero at $z = 3(1 - j)$. Assume the coefficient of the term z^4 to be unity. Find the Laurent series for $1/P(z)$ about $z = 1$.

59. Compute the residues of the following functions in their various poles:

$$\frac{5z^2 + z + 1}{z(z^2 + 3z + 8)} \qquad \frac{z^2 + 5e^{j2\pi/3}}{z^2 - 3e^{j\pi/6}} \qquad \frac{e^{jz}}{z}$$

60. A rational function $f(z)$ has simple poles at the points $z = z_1, z_2, \cdots z_n$ only, with the residues $k_1, k_2, \cdots k_n$ respectively. Find analytic expressions for the residues

of the function

$$\psi(z) = \frac{\zeta f(z)}{z(z - \zeta)}$$

in all its poles assuming that ζ does not coincide with any of the points $z_1 \cdots z_n$.

61. Continuing Prob. 60, evaluate the contour integral

$$\frac{1}{2\pi j} \oint_C \psi(z)\, dz$$

in which C is a circle about the origin with a radius R sufficiently large to enclose all the poles of $\psi(z)$. If $f(z)$ is regular at infinity, show that

$$\left| \oint_C \frac{f(z)\, dz}{z^2} \right| \to 0 \quad \text{as} \quad R \to \infty$$

and hence that the value of the above contour integral must be zero. Thus obtain the following form for the partial fraction expansion of $f(z)$:

$$f(\zeta) = f(0) + \sum_{\nu=1}^{n} k_\nu \left(\frac{1}{\zeta - z_\nu} + \frac{1}{z_\nu} \right)$$

62. Generalize the result of Prob. 61 to the extent of allowing $f(z)$ to possess an unlimited number of isolated simple poles (it is then no longer a rational function, of course). Through choosing the circular contour in such a way that no pole lies upon this path at any stage in the limiting process indicated by $R \to \infty$, conclude that the above partial fraction expansion of $f(z)$ yields (in the limit $n \to \infty$) a convergent series (theorem of Mittag-Leffler). Apply the result to verify the following expressions:

$$\frac{\pi}{\cos \pi z} = \pi + \sum_{\nu = -1, -3, -5 \cdots}^{1,3,5 \cdots} (-1)^n \left(\frac{1}{z - \frac{\nu}{2}} + \frac{1}{\frac{\nu}{2}} \right)$$

$$\pi \tan \pi z = - \sum_{\nu = -1, -3, -5 \cdots}^{1,3,5 \cdots} \left(\frac{1}{z - \frac{\nu}{2}} + \frac{1}{\frac{\nu}{2}} \right)$$

Convert these into summations over positive integers only.

63. If a rational function has a zero at infinity with a multiplicity of two or more, show that the sum of its residues is zero.

64. Find the values of the contour integrals

$$I_1 = \oint_C \frac{dz}{z^2 - 25} \qquad \text{and} \qquad I_2 = \oint_C \frac{z\, dz}{z^2 - 25}$$

when the contour C is defined as: (a) a circle of radius $R < 5$ with center at the origin, (b) a circle of radius $R > 5$ with center at the origin; (c) a triangle with vertices at the points $-6, j6, -j6$; (d) a triangle with vertices at the points $-6, j6, j10^{-6}$

65. Show that the function

$$f(z) = \frac{z}{z^3 + (-a + b + c)z^2 - (ab + ac - bc)z - abc} - \frac{a}{(a^2 + ab + ac + bc)(z - a)}$$

has a removable singularity at the point $z = a$ through demonstrating that $(z - a) \times f(z)$ is regular in this point. Compute the residues of $f(z)$ in the remaining poles.

66. Considering the function

$$f(z) = \frac{1}{\sin \frac{1}{z}}$$

within a circle of radius ρ about the origin, show that for a nonzero ρ however small, the circle nevertheless encloses an unlimited number of poles.

67. Starting from the definition of the Tschebyscheff polynomials

$$T_n(z) = \frac{\cos (n \cos^{-1} z)}{2^{n-1}}$$

derive the alternate relation

$$T_n(z) = \frac{(z + j\sqrt{1 - z^2})^n + (z - j\sqrt{1 - z^2})^n}{2^n}$$

and show that these functions are finite polynomials in z.

68. Find the points of stagnation of the function

$$f(z) = v_0\left(z + \frac{R^2}{z}\right) + j\frac{C}{2\pi} \ln z$$

in which v_0, R, and C are real constants, and map their locus in the z-plane as the quantity $C/(4\pi v_0 R)$ varies continuously from $-\infty$ through zero to ∞.

69. For any function $w = f(z) = u(x,y) + jv(x,y)$ satisfying the Cauchy-Riemann equations, regard $u(x,y)$ and $v(x,y)$ as representing altitude functions and thus defining a pair of surfaces s_1 and s_2. Show that a maximum or minimum in the altitude u coincides with a maximum or minimum in the altitude v and with a saddle point of the function $f(z)$.

70. Find the saddle points of the function

$$f(z) = e^{zt - a\sqrt{z^2 + b^2}}$$

in which t, a, and b are real constants. Show that they always lie in the real or imaginary axes and that their positions are controlled by the ratio t/a.

71. Inside a closed contour C the single-valued function $f(z)$ is regular and continuous except at poles in the points $a_1, a_2, \cdots a_n$ with the residues $b_1, b_2 \cdots b_n$. If L is the length of this contour and M is the maximum absolute value of $f(z)$ on the contour show that

$$\left|\sum_{k=1}^{n} b_k\right| \leq \frac{ML}{2\pi}$$

72. Consider the rational function

$$I(z) = \frac{a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m}{b_0 + b_1 z + b_2 z^2 + \cdots + b_n z^n}$$

and let $k_1, k_2, \cdots k_p \quad (p \leq n)$ be the residues in its poles which may or may not be simple. C is a circle about the origin with radius R large enough to enclose all the poles. Points on this circle are denoted by $z = Re^{j\phi}$. Show:

(a) That

$$2\pi|k_1 + k_2 + \cdots + k_p| \leqq R\int^{2\pi} |I(z)|\, d\phi$$

(b) If $n > (m + 1)$ that

$$(k_1 + k_2 + \cdots + k_p) = 0$$

(c) If $n = (m + 1)$ that

$$(k_1 + k_2 + \cdots + k_p) = \frac{a_m}{R^{n-m}b_n}$$

How can one obtain an expression for the sum of the residues when $n < (m + 1)$?

73. Evaluate the following integrals involving functions of a real variable x

$$\int_0^\infty \frac{x^2\, dx}{x^4 + 1} \quad \text{and} \quad \int_0^\infty \frac{dx}{x^6 + 1}$$

through replacing x by the complex variable $z = x + jy$ and using the methods of contour integration, choosing as the closed contour the real axis from $-R$ to R and a semicircle of radius R which is regarded as having unlimited magnitude Consider whether the semicircle should lie in the upper or lower half plane, and show that the contribution of the semicircular path increment to the value of the integral is negligible.

74. $z = e^{j\theta}$ is a point on the unit circle about the origin, and ψ (cos θ, sin θ) is a rational function of sin θ and cos θ Show that

$$\int_0^{2\pi} \psi\, d\theta = \oint_C f(z)\, dz$$

in which $f(z)$ is a rational function of z, and C is the unit circle about the origin.

75. Using the results of Prob. 74, verify the following integration

$$\int_0^{2\pi} \frac{d\theta}{1 + a^2 - 2a\cos\theta} = \frac{2\pi}{|a^2 - 1|}$$

76. Within a closed contour C the function $f(z)$ has simple poles at the points $z_1, z_2, \cdots z_n$ with the residues $b_1, b_2, \cdots b_n$ respectively. Inside the same contour the function cot πz has p simple poles. Show that

$$\frac{1}{2j}\oint_C f(z)\cot\pi z\, dz = \sum_{k=1}^{p} f(m + k) + \pi\sum_{k=1}^{n} b_k \cot\pi z_k$$

in which m is an appropriate integer.

77. Through replacing x by z, choosing a closed contour consisting of one large semicircle of radius R, a small one of radius ρ (both in the upper or lower half plane) confluent with portions of the real axis, and showing that contributions due to integration along the semicircular path increments vanish in the limits $R \to \infty$ and $\rho \to 0$, obtain the result

$$\int_{-\infty}^{\infty} \frac{(\ln x)^2}{1 + x^2}\, dx = -\frac{\pi^3}{4}$$

It is now proposed to find the integral of the same function between the limits 0 and ∞ through considering the change of variable $x \to (-x)$ in the integration over $-\infty$ to 0,

noting that

$$[\ln(-x)]^2 = (\ln x \pm j\pi)^2 = (\ln x)^2 - \pi^2 \pm j2\pi \ln x$$

Thus, after obtaining the collateral results,

$$\int_0^\infty \frac{dx}{1+x^2} = \frac{\pi}{2} \quad \text{and} \quad \int_0^\infty \frac{\ln x}{1+x^2}\,dx = 0$$

show that

$$\int_0^\infty \frac{(\ln x)^2}{1+x^2}\,dx = \frac{\pi^3}{8}$$

78. By means of complex integration, following a pattern similar to that suggested in Prob. 77, check the values of the following definite integrals

$$\int_0^\infty \frac{\cos x}{x^2+a^2}\,dx = \frac{\pi e^{-a}}{2a} \qquad \int_0^\infty \frac{\sin x}{x^2+a^2}\,dx = \frac{\pi e^{-a}}{2a}\sinh a$$

Express $\cos z$ and $\sin z$ in terms of the exponential function and correspondingly represent each integral as a sum of components. Choose composite contours consisting of the real axis and a semicircle of radius $R \to \infty$, placing the latter in the upper or lower half plane according to the requirement that this path increment shall contribute nothing to the value of the contour integral. Note that the second of the two integrals above remains proper for $a \to 0$ If the constant a is regarded as real throughout the process of evaluation, is the result nevertheless still valid for any complex a-value, for example, for a pure imaginary value?

79. If $f(z)$ is a rational function having a zero at infinity with a multiplicity equal to or greater than one, and if $F(t)$ is a real function of the real variable t which is zero for $t < 0$ and has the property that $F(t)e^{-zt} \to 0$ for $t \to \infty$ (assuming z to have a positive real part), the following pair of mutual integral relations hold:

$$F(t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} f(z)e^{zt}\,dz \qquad f(z) = \int_{-\infty}^{\infty} F(t)e^{-zt}\,dt$$

Check these relations for the following pairs of functions:

$$f(z) = \frac{n!}{(z+a)^{n+1}} \qquad F(t) = t^n e^{-at}$$

$$f(z) = \frac{a}{z^2+a^2} \qquad F(t) = \sin at$$

$$f(z) = \frac{z}{z^2+a^2} \qquad F(t) = \cos at$$

$$f(z) = \frac{z+a}{(z+a)^2+b^2} \qquad F(t) = e^{-at}\cos bt$$

The first of the two integrals is to be evaluated using complex integration along a closed contour consisting of the imaginary axis and a large semicircle of radius $R \to \infty$ lying in the left or right half plane according to whether $t > 0$ or $t < 0$ (see Art. 26 of Ch. VII). The second integral in the above pair is to be evaluated according to the methods of real integration.

80. Expand the following rational functions in partial fractions and check the

relations regarding the residues as stated in parts (*b*) and (*c*) of Prob. 72:

$$\text{(a)}\ \frac{5z^4(z^2+2z-1)}{(z^2-2)(z^2-z-1)} \qquad \text{(b)}\ \frac{z(z^2+2z-1)}{(z-3)^2(z+8)}$$

$$\text{(c)}\ \frac{z^2+2z-1}{(z-3)^2(z+8)} \qquad \text{(d)}\ \frac{1}{(z-2)^2(z-3)^3}$$

81. Discuss the structure of the Riemann surface associated with each of the following functions:

$$\text{(a)}\ w = \sqrt[3]{z} \qquad \text{(b)}\ \sqrt{\frac{2z-3}{z}}$$

$$\text{(c)}\ w = \sqrt{(z-a)(z-b)^2} \qquad \text{(d)}\ w = \sqrt{(z-5)(z^2+6z+13)}$$

pointing out the location and order of branch points and other singularities as well as the character of the point $z = \infty$.

82. Discuss the Riemann surfaces in both the w- and z-planes associated with the function $w = z^{3/5}$.

83. For the function $w = \sqrt[3]{z}$ consider the leaves of the Riemann surface defined by the statements:

$-\pi <$ arg. $z < \pi$ corresponds to leaf I.

$\pi <$ arg. $z < 3\pi$ corresponds to leaf II.

$3\pi <$ arg. $z < 5\pi$ corresponds to leaf III.

In the w-plane plot loci corresponding to the following straight lines in the z-plane:

(a) Parallel to the real axis at distances 0.1 and -0.1 from the origin and lying in leaf I.

(b) Parallel to the imaginary axis at the distance 0.1 from the origin and lying in leaf III.

(c) Parallel to the imaginary axis at the distance -0.1 from the origin and lying in leaves I and II.

84. For each of the following functions $w = f(z)$ construct the algebraic equation $F(z,w) = 0$, which generates the complete function, and from this equation determine the inverse function $z = \phi(w)$.

$$\text{(a)}\ w = \sqrt{z^2-1} \qquad \text{(b)}\ w = z + \sqrt{z^2+1} \qquad \text{(c)}\ w = \sqrt{\frac{1+z^2}{1-z^2}}$$

Describe in each case the structure of the Riemann surfaces in the w- and z-planes, pointing out the locations and multiplicities of branch points, branch cuts, etc.

85. Recognizing that each of the complete functions $w = \sqrt{z}$ and $W = \sqrt{z-1}$ has two branches, construct the algebraic equation that generates the function $U = w + W$. Determine this function as well as its inverse $z = \phi(U)$, and discuss completely the Riemann surfaces in the U- and z-planes.

86. The expressions

$$w = \sqrt{z^2} \quad \text{and} \quad w = \sqrt{1 - \sin^2 z}$$

represent pairs of single-valued functions rather than multivalued functions. Demonstrate the truth of this statement through showing that the two functions of a pair cannot be obtained one from the other by the process of analytic continuation.

87. Through the use of Hilbert transforms solve the following:

(a) given $u(0,y) = \begin{cases} 1 & \text{for } -y_1 < y < y_1 \\ 0 & \text{for } |y| > y_1 \end{cases}$ find $v(0,y)$

(b) given $u(0,y) = \begin{cases} 0 & \text{for } -y_1 < y < y_1 \\ a & \text{for } |y| > y_1 \end{cases}$ find $v(0,y)$

(c) given $u(0,y) = \begin{cases} 1 + \dfrac{y}{y_1} & \text{for } -y_1 < y < 0 \\ 1 - \dfrac{y}{y_1} & \text{for } 0 < y < y_1 \\ 0 & \text{for } |y| > y_1 \end{cases}$ find $v(0,y)$

(d) given $u(0,y) = \begin{cases} -\dfrac{ay}{y_1} & \text{for } -y_1 < y < 0 \\ \dfrac{ay}{y_1} & \text{for } 0 < y < y_1 \\ a & \text{for } |y| > y_1 \end{cases}$ find $v(0,y)$

(e) given $u(0,y) = \begin{cases} \frac{1}{2} + \dfrac{(y + y_1)}{\delta} & \text{for } -y_1 - \dfrac{\delta}{2} < y < -y_1 + \dfrac{\delta}{2} \\ \frac{1}{2} - \dfrac{(y - y_1)}{\delta} & \text{for } y_1 - \dfrac{\delta}{2} < y < y_1 + \dfrac{\delta}{2} \\ 1 & \text{for } |y| < y_1 - \dfrac{\delta}{2} \\ 0 & \text{for } |y| > y_1 + \dfrac{\delta}{2} \end{cases}$ find $v(0,y)$

(f) given $v(0,y) = \begin{cases} \dfrac{n\pi}{y_1} y & \text{for } -y_1 < y < y_1 \\ -n\pi & \text{for } y < -y_1 \\ n\pi & \text{for } y > y_1 \end{cases}$ find $u(0,y)$

CHAPTER VII

Fourier Series and Integrals

1. Finite trigonometric polynomials

In discussing the convergence of Fourier series it is necessary to have compact expressions for their partial sums. For this reason, and also because one finds expressions of this sort useful in various other problems having to do with trigonometric series, a number of formulas are developed whereby a variety of finite trigonometric polynomials are given in closed form.

In view of the well-known geometric series

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \cdots \qquad [1]$$

and

$$\frac{z^{n+1}}{1-z} = z^{n+1} + z^{n+2} + z^{n+3} + \cdots \qquad [2]$$

one may write

$$\frac{1-z^{n+1}}{1-z} = 1 + z + z^2 + \cdots + z^n \qquad [3]$$

and

$$\frac{1}{z^n}\frac{1-z^{n+1}}{1-z} = 1 + z^{-1} + z^{-2} + \cdots + z^{-n} \qquad [4]$$

By addition and subtraction respectively these equations yield

$$\frac{1-z^{n+1}}{1-z}(1+z^{-n}) = 2\{1 + \tfrac{1}{2}(z+z^{-1}) + \tfrac{1}{2}(z^2+z^{-2}) + \cdots + \tfrac{1}{2}(z^n+z^{-n})\} \qquad [5]$$

and

$$\frac{1-z^{n+1}}{1-z}(1-z^{-n}) = 2\{\tfrac{1}{2}(z-z^{-1}) + \tfrac{1}{2}(z^2-z^{-2}) + \cdots + \tfrac{1}{2}(z^n-z^{-n})\} \qquad [6]$$

Letting

$$z = e^{jx} = \cos x + j \sin x \qquad [7]$$

and noting that

$$z^n = e^{jnx} = \cos nx + j \sin nx \qquad [8]$$

one finds for example, that

$$\frac{1-z^{n+1}}{1-z}(1+z^{-n}) = \frac{z^{(n+1)/2}-z^{-(n+1)/2}}{z^{1/2}-z^{-1/2}}(z^{n/2}+z^{-n/2})$$

$$= \frac{2\sin(n+1)\frac{x}{2}\cos n\frac{x}{2}}{\sin\frac{x}{2}} \qquad [9]$$

By use of well-known trigonometric identities, this may further be transformed as shown by

$$\frac{2\sin(n+1)\frac{x}{2}\cos n\frac{x}{2}}{\sin\frac{x}{2}} = \frac{\sin(2n+1)\frac{x}{2}}{\sin\frac{x}{2}} + 1 \qquad [10]$$

The use of manipulations of this sort enables one to obtain from Eqs. 5 and 6 the formulas

$$\frac{1}{2} + \frac{\sin(2n+1)\frac{x}{2}}{2\sin\frac{x}{2}} = 1 + \cos x + \cos 2x + \cdots + \cos nx \qquad [11]$$

and

$$\frac{1}{2}\cot\frac{x}{2} - \frac{\cos(2n+1)\frac{x}{2}}{2\sin\frac{x}{2}} = \sin x + \sin 2x + \cdots + \sin nx \qquad [12]$$

The compact forms thus obtained for the finite trigonometric polynomials given by the right-hand sides of Eqs. 11 and 12 are useful in a variety of practical problems as well as in connection with various theoretical discussions regarding Fourier series. Other formulas are readily obtained from Eqs. 11 and 12. For example, replacing the variable x in Eq. 11 by $(x + \pi)$ yields

$$\frac{1}{2} + \frac{(-1)^n\cos(2n+1)\frac{x}{2}}{2\cos\frac{x}{2}} = 1 - \cos x + \cos 2x - \cos 3x + \cdots + (-1)^n\cos nx \qquad [13]$$

which is the same polynomial but with alternating signs. Making the same

change of variable in Eq. 12 gives

$$\frac{1}{2}\tan\frac{x}{2} + \frac{(-1)^{n-1}\sin(2n+1)\frac{x}{2}}{2\cos\frac{x}{2}} = \sin x - \sin 2x + \sin 3x - \cdots + (-1)^{n-1}\sin nx \quad [14]$$

Subtracting Eq. 13 for n odd from Eq. 11 yields a formula for a cosine polynomial with only odd integer multiples of x, thus:

$$\frac{\sin(n+1)x}{2\sin x} = \cos x + \cos 3x + \cos 5x + \cdots + \cos nx \quad [15]$$

Similarly, adding Eq. 12 and Eq. 14 for n odd yields

$$\frac{1-\cos(n+1)x}{2\sin x} = \sin x + \sin 3x + \sin 5x + \cdots + \sin nx \quad [16]$$

in which $\sin^2(n+1)\frac{x}{2}$ may be written instead of $\frac{1}{2}[1-\cos(n+1)x]$ if desired.

In Eq. 15 one may introduce the change of variable $x \to x + \frac{\pi}{2}$ and obtain a sine polynomial with odd multiples of x and alternating signs, thus:

$$\frac{(-1)^{(n-1)/2}\sin(n+1)x}{2\cos x} = \sin x - \sin 3x + \sin 5x - \cdots + (-1)^{(n-1)/2}\sin nx \quad [17]$$

Making the same change of variable in Eq. 16 yields

$$\frac{1+(-1)^{(n-1)/2}\cos(n+1)x}{2\cos x} = \cos x - \cos 3x + \cos 5x - \cdots + (-1)^{(n-1)/2}\cos nx \quad [18]$$

In these last two transformations it is significant to note that Eqs. 15 and 16 apply only to odd integer values of n, so that $(n+1)$ is an even integer. Hence $\sin(n+1)\frac{\pi}{2} = 0$; whence

$$\cos(n+1)\left(x+\frac{\pi}{2}\right) = \cos(n+1)x\cos(n+1)\frac{\pi}{2} - \sin(n+1)x\sin(n+1)\frac{\pi}{2}$$
$$= (-1)^{(n+1)/2}\cos(n+1)x = -(-1)^{(n-1)/2}\cos(n+1)x \quad [19]$$

and

$$\sin (n+1)\left(x+\frac{\pi}{2}\right) = \sin (n+1)x \cos (n+1)\frac{\pi}{2} + \cos (n+1)x \sin (n+1)\frac{\pi}{2}$$
$$= (-1)^{(n+1)/2} \sin (n+1)x = -(-1)^{(n-1)/2} \sin (n+1)x \quad [20]$$

2. The orthogonality relations and their significance in the expansion of arbitrary functions

The trigonometric functions, in common with many other kinds of so-called systems of *proper functions*, possess a very interesting and important property which greatly facilitates the process of representing an arbitrary function over a given interval by a series in terms of proper functions. Since the underlying principles of this process have a broad significance in the solution to problems in potential theory and wave motion, a few rather general introductory remarks are in order.

For detailed discussion of the derivation of the following fundamental differential equations and for physical interpretation of them, the reader is referred to other portions of this reference series, the chief interest at the moment being focused primarily upon their purely mathematical import. In dealing with problems in potential theory, in which a desired potential function is a function of the space co-ordinates (x,y,z) alone, Laplace's equation

$$\nabla^2 V = 0 \quad [21]$$

is found to govern the behavior of that function, whereas in problems involving wave motion, in which the time co-ordinate also is involved, this equation is modified by the appearance of an additional term so that it reads

$$\nabla^2 \mathfrak{V} - \frac{1}{c^2}\frac{\partial^2 \mathfrak{V}}{\partial t^2} = 0 \quad [22]$$

In the first of these two cases the potential is static, whereas in the second it is dynamic — that is, it varies with the time as well as with the space co-ordinates. In other words, a problem in wave motion is simply a problem in potential theory with an added time co-ordinate.

The form of the Laplacian operator ∇^2 depends upon the particular system of space co-ordinates used.* In the commonest of these, the

*See Art. 18, Ch. V.

ordinary rectangular Cartesian system, it is given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad [23]$$

In dealing with the so-called *wave equation* 22, the first step is to eliminate the time variable. This is done through assuming

$$\mathcal{V} = Ve^{j\omega_n t} \qquad [24]$$

whence

$$\frac{\partial^2 \mathcal{V}}{\partial t^2} = -\omega_n^2 \mathcal{V} \qquad [25]$$

so that Eq. 22, after cancellation of the common factor $e^{j\omega_n t}$, reads

$$\nabla^2 V + k_n^2 V = 0 \qquad [26]$$

in which*

$$k_n^2 = \frac{\omega_n^2}{c^2} \qquad [27]$$

In Eq. 26, as in Eq. 21, the function V is a function of the space co-ordinates only.

The particular form of these equations now depends upon the system of co-ordinates used, and the choice of co-ordinates in turn depends upon the geometry of the physical system to which the equations apply. For a rectangular geometry the ordinary Cartesian co-ordinates are used; for a cylindrical geometry, cylindrical co-ordinates; for a spherical geometry, spherical co-ordinates; and so on. Correspondingly, these particular forms of the equations are known by certain names, such as Bessel's equation (for cylindrical co-ordinates) or Legendre's equation (for spherical co-ordinates), and the particular types of functions which formally satisfy the equations are largely known by names which relate to the particular geometry of the physical system, such as *cylinder functions* (of which the Bessel functions are the so-called *first kind*) or spherical harmonics (also known as Legendre polynomials), and so on. These functions are referred to as the *proper* functions pertaining to the particular physical system under consideration, the simplest of them being the trigonometric functions which are the proper functions for systems having a rectangular geometry. In a single dimension this type of geometry is that of a straight line, like a stretched violin string, for example.

In terms of some parameter, like parameter n in sin nx and cos nx

*The values $k_1, k_2, \cdots$, which usually are infinite in number, are determined from conditions which the function $\mathcal{V}(x,y,z,t)$ is required to satisfy at certain physical boundaries. The details of this process need not be considered at the moment.

or, in the case of solutions to the wave equation, in terms of the so-called *proper values* k_n, these functions form a set or system. In view of the linearity of the equations, it follows that a complete formal solution is given by a linear superposition of a set of these proper functions with different parameter values and arbitrary coefficients. Thus if $\phi_n(x,y,z)$ represents a proper function for the parameter index n, the solution has the form

$$V = a_1\phi_1 + a_2\phi_2 + a_3\phi_3 + \cdots \qquad [28]$$

which in general is an infinite series. The coefficients a_n are regarded as constants of integration which give the formal solution 28 the necessary flexibility of meeting certain boundary conditions set by the physical problem.

Thus in a two-dimensional problem in static potential theory, for example, the potential $\mathcal{V}(x,y)$ may have to become identical with a certain function $f(x)$ for y equal to a particular value, say $y = 0$, which characterizes a physical boundary. Or, in a problem of wave motion, the function $\mathcal{V}(x,y,z,t)$ may, for the temporal boundary $t = 0$, for which (according to Eq. 24) $\mathcal{V} = V(x,y,z)$, have to meet some prescribed function. For example, a stretched membrane for which $\mathcal{V}$ represents the displacement of various points from an equilibrium position is given a particular deformation from which it is suddenly released at an instant which is designated as $t = 0$, or an electrical transmission line (which involves a single space co-ordinate) has, at a given initial instant, distributed upon it certain charges which give rise to a specific potential function versus distance along the line. These are referred to as *initial* or *boundary distributions*. Inasmuch as they may, in a physical system, be arbitrarily specified, the process of solution must be able to fit a series of the form given by Eq. 28 to a specified function of one or more of the independent variables. The fulfillment of these conditions may necessitate the selection of certain kinds of proper functions any of which formally satisfy the differential equations, such as the selection, for example, of a particular kind of cylinder function in the solution to a problem with cylindrical symmetry.

In one dimension this problem takes the form

$$f(x) = a_1\phi_1(x) + a_2\phi_2(x) + a_3\phi_3(x) + \cdots \qquad [29]$$

in which $f(x)$ and the functions $\phi_n(x)$ are known but the coefficients a_n are to be determined so as to satisfy this equation. The problem here presented is determining the expansion of an arbitrary function $f(x)$ in a series of specified proper functions in such a way that the resulting series is in general a convergent one.

The solution to this problem is either impossible, or possible only

through the use of ingenious artifices, unless the system of proper functions (or a derived system formed from linear combinations of these functions) satisfies the so-called *conditions of orthogonality*, which in the one-dimensional case are expressed by the equations

$$\int_a^b \phi_m(x)\phi_n(x)\,dx = \begin{cases} r_n & \text{for } m = n \\ 0 & \text{for } m \neq n \end{cases} \qquad [30]$$

in which a and b are the finite limits of the region over which the function $f(x)$ is specified. The quantity r_n is ordinarily a constant and can be made equal to unity by the incorporation of an appropriate scale factor. This process is called *normalization*, and the resulting functions $\phi_n(x)$ are then spoken of as a *normalized set of orthogonal proper functions* or an *orthonormal set.*

The use of the term " orthogonality " in connection with the relations 30 is suggested by the parallelism between these conditions and those characterizing an orthogonal matrix.* Since the relationship between the present problem and that presented by a set of linear algebraic equations, which is thus implied, has rather more than superficial significance, this item is discussed in greater detail.

Regarding the partial sum

$$s_n(x) = a_1\phi_1(x) + a_2\phi_2(x) + \cdots + a_n\phi_n(x) \qquad [31]$$

corresponding to the series 29, and assuming that the function $s_n(x)$ is specified over the range $a \leqq x \leqq b$, one finds that a possible procedure for determining the coefficients a_k such that the finite polynomial in terms of the functions $\phi_k(x)$ approximates the given $s_n(x)$ is to divide the interval a to b into n equal subintervals whose midpoints are the x-values expressed by $a < x_1 < x_2 < \cdots < x_n < b$, and then to write the set of n equations

$$\begin{aligned} s_n(x_1) &= a_1\phi_1(x_1) + a_2\phi_2(x_1) + \cdots + a_n\phi_n(x_1) \\ s_n(x_2) &= a_1\phi_1(x_2) + a_2\phi_2(x_2) + \cdots + a_n\phi_n(x_2) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ s_n(x_n) &= a_1\phi_1(x_n) + a_2\phi_2(x_n) + \cdots + a_n\phi_n(x_n) \end{aligned} \qquad [32]$$

in which the n coefficients a_k are regarded as unknowns. The values of the coefficients thus determined, when substituted into Eq. 31, yield a function $s_n(x)$ which has the correct values at the selected points $x_1 \cdots x_n$ within the prescribed interval. At any other points, not much can be said about the degree of approximation afforded by the finite sum in Eq. 31. However, by taking n sufficiently large, and assuming for the moment that $s_n(x)$ is a smooth function, one may expect a solution which fulfills the requirements of a given physical problem. Indeed, as n is

*See Ch. II, Art. 6, and Ch. III, Art. 4.

chosen larger and larger, one may expect a closer and closer approximation to the partial sum $s_n(x)$, which is ultimately identified with the specified function $f(x)$.

Such a process of solution requires, for a finite n, the inversion of the matrix

$$\begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_n(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_1(x_n) & \phi_2(x_n) & \cdots & \phi_n(x_n) \end{bmatrix} \qquad [33]$$

which in general is a laborious task, and becomes hopeless if not impossible as n is increased without limit. However, if this matrix is an orthogonal one, its inverse is simply given by the transpose of the matrix 33. The solution for any n is then immediately written down.

The orthogonality conditions for the matrix 33 read*

$$\sum_{k=1}^{n} \phi_r(x_k)\phi_s(x_k) = \begin{cases} 1 & \text{for } r = s \\ 0 & \text{for } r \neq s \end{cases} \qquad [34]$$

which are readily recognized as having the same form as the relations 30. The solutions to Eqs. 32 are then given by

$$a_s = \sum_{k=1}^{n} \phi_s(x_k)s_n(x_k) \qquad [35]$$

This result may be obtained through multiplying Eqs. 32 successively by $\phi_s(x_1)$, $\phi_s(x_2)$, $\cdots \phi_s(x_n)$, and adding. In view of the conditions 34, the sums of all the columns vanish except that of the sth, which yields a_s, whereas the sum of the left-hand members is seen to be given by the right-hand side of Eq. 35.

The equivalent solution for the coefficients in the infinite series 29 is obtained through multiplying this series by $\phi_s(x)$ and integrating over the region $a \leqq x \leqq b$. Then if the conditions 30 hold with $r_n = 1$, there results

$$a_s = \int_a^b \phi_s(x)f(x)\,dx \qquad [36]$$

which represents the desired solution.

In order to derive the orthogonality conditions for the trigonometric

*These are the orthogonality conditions for the columns only. The corresponding ones for the rows are not needed for the present discussion. Indeed, it is actually not necessary for the matrix to be orthogonal since the existence of orthogonality conditions for only its columns is sufficient.

functions it is convenient to write them in the exponential form

$$\sin nx = \frac{1}{2j}(e^{jnx} - e^{-jnx}) \qquad [37]$$

$$\cos nx = \frac{1}{2}(e^{jnx} + e^{-jnx}) \qquad [38]$$

whence

$$\sin mx \sin nx = -\frac{1}{4}(e^{j(m+n)x} + e^{-j(m+n)x} - e^{j(m-n)x} - e^{-j(m-n)x}) \qquad [39]$$

$$\sin mx \cos nx = \frac{1}{4j}(e^{j(m+n)x} - e^{-j(m+n)x} + e^{j(m-n)x} - e^{-j(m-n)x}) \qquad [40]$$

and

$$\cos mx \cos nx = \frac{1}{4}(e^{j(m+n)x} + e^{-j(m+n)x} + e^{j(m-n)x} + e^{-j(m-n)x}) \qquad [41]$$

Next it is observed that

$$\int_a^{a+2\pi} e^{jkx}\,dx = \frac{e^{jkx}}{jk}\bigg]_a^{a+2\pi} = \frac{e^{jka}(e^{jk2\pi} - 1)}{jk} \qquad [42]$$

Since $e^{jk2\pi} = 1$ for all integer values of k, the numerator in this result is zero for all k-values. The denominator is not zero except for $k = 0$. Hence the value of the integral in Eq. 42 vanishes for all k-values except $k = 0$, when the right-hand side of this equation assumes an indeterminate form. The latter is readily evaluated if the limit $k \to 0$ is considered to proceed gradually. For small values of k the exponential $e^{jk2\pi}$ may be replaced by a few terms of its Maclaurin series, giving

$$\int_a^{a+2\pi} e^{jkx}\,dx \to \frac{e^{jka}(1 + jk2\pi + \cdots - 1)}{jk} \to 2\pi \qquad [43]$$

Thus it becomes clear that

$$\int_a^{a+2\pi} e^{jkx}\,dx = \begin{cases} 2\pi & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases} \qquad [44]$$

This is an important property of the exponential function.

By means of the expressions 39, 40, and 41 it is now readily seen that

$$\int_a^{a+2\pi} \sin mx \sin nx\,dx = \begin{cases} \pm\pi & \text{for } m = \pm n \neq 0 \\ 0 & \text{for } m \neq \pm n \end{cases} \qquad [45]$$

$$\int_a^{a+2\pi} \sin mx \cos nx\,dx = 0 \quad \text{for all } m \text{ and } n \qquad [46]$$

$$\int_a^{a+2\pi} \cos mx \cos nx\,dx = \begin{cases} +\pi & \text{for } m = \pm n \neq 0 \\ 0 & \text{for } m \neq \pm n \end{cases} \qquad [47]$$

The *fundamental range*, so-called, which in the preceding discussion of this article is indicated by $a \leqq x \leqq b$, is in the integrals 44, 45, 46, and 47 seen to be any region throughout which the variable x changes by an increment of 2π. The limits on these integrals are, therefore, arbitrary except that they must differ by 2π (or a multiple of 2π).

By the same process of analysis it is also readily found that

$$\int_{\nu\pi}^{(\nu+1)\pi} e^{jkx}\,dx = \begin{cases} \dfrac{2(-1)^{\nu-1}}{jk} & \text{for } k \text{ odd} \\ 0 & \text{for } k \text{ even except} \\ \pi & \text{for } k = 0 \end{cases} \qquad [48]$$

in which ν is any integer. In other words, if the fundamental range has a "width" of only π instead of 2π, a result similar to the one expressed by Eq. 44 exists only for k even, and then only when the lower limit on the integral is an integer multiple (including zero) of π.

If it is now observed that, for any integers m and n, $(m + n)$ and $(m - n)$ are either both even or both odd, the result 48 together with the relations 39, and 41 shows that

$$\int_{\nu\pi}^{(\nu+1)\pi} \sin mx \sin nx\,dx = \begin{cases} \pm\dfrac{\pi}{2} & \text{for } m = \pm n \neq 0 \\ 0 & \text{for } m \neq \pm n \end{cases} \qquad [49]$$

and

$$\int_{\nu\pi}^{(\nu+1)\pi} \cos mx \cos nx\,dx = \begin{cases} +\dfrac{\pi}{2} & \text{for } m = \pm n \neq 0 \\ 0 & \text{for } m \neq \pm n \end{cases} \qquad [50]$$

in which ν is any integer including zero. In other words, the trigonometric functions $\sin nx$ and $\cos nx$ also form orthogonal sets over a fundamental range which is only π units wide. However, it is to be observed that a relation similar to the one expressed by Eq. 46 does not hold for a range of this width.

If the range $0 < x < \pi$ is divided into n equal subranges, and the values $x_1, x_2, \cdots x_n$ refer to the centers of these subranges, as shown in Figs. 1 and 2 for $n = 8$, then matrices corresponding to the matrix 33 may be formed for the sine and cosine functions.

Thus it is found that

$$x_k = \frac{(2k-1)\pi}{2n} \qquad [51]$$

so that if

$$\alpha_{ks} = \cos s x_k \qquad [52]$$

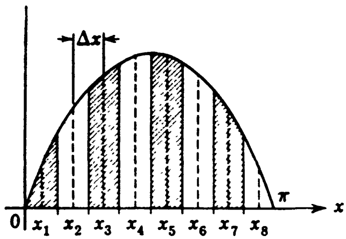


FIG. 1. Division of sine into subranges appropriate to the formation of an orthogonal matrix.

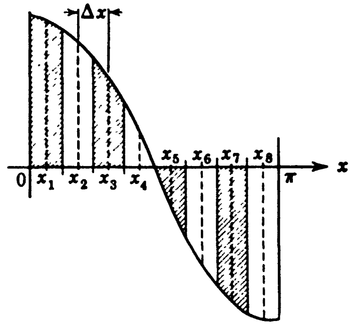


FIG. 2. Division of cosine into subranges appropriate to the formation of an orthogonal matrix.

and

$$\beta_{ks} = \sin sx_k \qquad [53]$$

then the matrices

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix} \qquad [54]$$

and

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nn} \end{bmatrix} \qquad [55]$$

satisfy the conditions

$$\sum_{k=1}^{n} \alpha_{kr}\alpha_{ks} = \begin{cases} \dfrac{n}{2} & \text{for } r = \pm s \quad \text{except} \\ n & \text{for } r = s = 0 \\ 0 & \text{for } r \neq \pm s \quad \text{or } r = \pm(s = n) \end{cases} \qquad [56]$$

and

$$\sum_{k=1}^{n} \beta_{kr}\beta_{ks} = \begin{cases} \pm\dfrac{n}{2} & \text{for } r = \pm s \quad \text{except} \\ \pm n & \text{for } r = \pm(s = n) \\ 0 & \text{for } r \neq \pm s \quad \text{or } r = s = 0 \end{cases} \qquad [57]$$

These may be put into a form in which they more nearly simulate the expressions 49 and 50 through noting, with reference to Figs. 1 and 2, that

$$\Delta x = \frac{\pi}{n} \tag{58}$$

Hence

$$\sum_{k=1}^{n} \alpha_{kr}\alpha_{ks}\,\Delta x = \begin{cases} \dfrac{\pi}{2} & \text{for } r = \pm s \quad \text{except} \\ \pi & \text{for } r = s = 0 \\ 0 & \text{for } r \neq \pm s \quad \text{or } r = \pm(s = n) \end{cases} \tag{59}$$

and

$$\sum_{k=1}^{n} \beta_{kr}\beta_{ks}\,\Delta x = \begin{cases} \pm\dfrac{\pi}{2} & \text{for } r = \pm s \quad \text{except} \\ \pm\pi & \text{for } r = \pm(s = n) \\ 0 & \text{for } r \neq \pm s \quad \text{or } r = s = 0 \end{cases} \tag{60}$$

These, or the expressions 56 and 57, may be verified independently. Thus

$$\alpha_{kr}\alpha_{ks} = \cos rx_k \cos sx_k = \tfrac{1}{2}\{\cos(r-s)x_k + \cos(r+s)x_k\} \tag{61}$$

and

$$\beta_{kr}\beta_{ks} = \sin rx_k \sin sx_k = \tfrac{1}{2}\{\cos(r-s)x_k - \cos(r+s)x_k\} \tag{62}$$

Now by Eq. 51

$$\sum_{k=1}^{n} \cos(r-s)x_k = \sum_{k=1}^{n} \cos\left\{\frac{(2k-1)(r-s)\pi}{2n}\right\}$$

$$= \cos\frac{(r-s)\pi}{2n} + \cos\frac{3(r-s)\pi}{2n} + \cdots + \cos\frac{(2n-1)(r-s)\pi}{2n} \tag{63}$$

This expression is the same as the right-hand side of Eq. 15 of the previous article except that $\dfrac{(r-s)\pi}{2n}$ takes the place of x, and $(2n-1)$ takes the place of n. Hence it is found that

$$\sum_{k=1}^{n} \cos(r-s)x_k = \frac{\sin(r-s)\pi}{2\sin\dfrac{(r-s)\pi}{2n}} = \begin{cases} n & \text{for } r = s \\ 0 & \text{for } r \neq s \quad \text{except} \\ -n & \text{for } r = -(s = n) \end{cases} \tag{64}$$

Similarly

$$\sum_{k=1}^{n} \cos(r+s)x_k = \frac{\sin(r+s)\pi}{2\sin\dfrac{(r+s)\pi}{2n}} = \begin{cases} n & \text{for } r = -s \\ 0 & \text{for } r \neq -s \quad \text{except} \\ -n & \text{for } r = (s = n) \end{cases} \tag{65}$$

The last two relations, together with those stated by Eqs. 61 and 62, verify the conditions given by Eqs. 56 and 57.

If the functions α_{ks} and β_{ks} are normalized through division by $\sqrt{n/2}$, it is observed that they fulfill conditions of the form given by Eq. 34 except for $r = s = n$. This means that if the coefficients in the partial sum 31 are approximately evaluated by means of the formula 35, a correction must be applied for the coefficient a_n. It should be observed in this connection, however, that the values of the coefficients a_k obtained by such a method become more and more approximate the nearer the index k approaches n. As pointed out previously, the partial sum $s_n(x)$ converges toward the desired function $f(x)$ only as n is increased without limit. For a sufficiently large finite n, the partial sum $s_n(x)$ may be regarded as an approximation to $f(x)$ which is close enough for certain practical purposes, but the degree of the approximation is then very likely little affected if one entirely ignores the last term, or perhaps several terms in this vicinity of the partial sum $s_n(x)$. This item is further discussed in Art. 15.

It may also be observed that the matrices 54 and 55, with the elements as defined by Eqs. 52 and 53, are actually not orthogonal matrices since the orthogonality conditions are fulfilled only for their columns, and not for their rows. As far as the present problem is concerned, however, this situation is not significant since none of the reasoning given in the present article is thereby affected.

3. The Fourier Series

The previous article indicates how a representation in the form of an infinite trigonometric series may be found for a given function $f(x)$ specified over a finite interval $a \leqq x \leqq a + 2\pi$. Incidentally, this designation for the interval may be regarded as perfectly general inasmuch as any other, such as $\alpha < x < \alpha + \ell$, is converted to it by means of the change of variable $x \to \frac{x\ell}{2\pi}$.

The trigonometric series, or *Fourier series*, as it is called, has the form

$$f(x) = \frac{a_0}{2} + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + \cdots \\ + b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots \qquad [66]$$

The necessity of having both sine and cosine terms is readily recognized. Writing the series 66 in the more compact form

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx + \sum_{n=1}^{\infty} b_n \sin nx \qquad [67]$$

multiplying both sides by cos mx, and integrating over the fundamental range, one has

$$\int_a^{a+2\pi} f(x)\cos mx\,dx = \frac{a_0}{2}\int_a^{a+2\pi}\cos mx\,dx + \int_a^{a+2\pi}\sum_{n=1}^{\infty} a_n\cos nx\cos mx\,dx + \int_a^{a+2\pi}\sum_{n=1}^{\infty} b_n\sin nx\cos mx\,dx \tag{68}$$

The first integral on the right-hand side is zero except for $m = 0$. Excluding this value, and assuming for the moment that the series 66 or 67 converges uniformly over the fundamental range $a \leqq x \leqq a + 2\pi$, so that the integration of the infinite sums may be carried out term by term, one finds

$$\int_a^{a+2\pi} f(x)\cos mx\,dx = \sum_{n=1}^{\infty} a_n\int_a^{a+2\pi}\cos nx\cos mx\,dx + \sum_{n=1}^{\infty} b_n\int_a^{a+2\pi}\sin nx\cos mx\,dx \tag{69}$$

In view of the conditions expressed by Eqs. 46 and 47, one obtains

$$a_m = \frac{1}{\pi}\int_a^{a+2\pi} f(x)\cos mx\,dx \tag{70}$$

This formula evaluates the coefficients of the cosine series, and incidentally gives the coefficient a_0 correctly also, inasmuch as the orthogonality conditions 46 and 47 applied to the term-by-term integration in Eq. 68 yields for $m = 0$

$$\int_a^{a+2\pi} f(x)\,dx = \frac{a_0}{2}\int_a^{a+2\pi} dx = a_0\pi \tag{71}$$

If Eq. 67 is multiplied by sin mx and integrated term by term over the fundamental range, it being recognized that the term with a_0 then yields zero even for $m = 0$, one obtains

$$\int_a^{a+2\pi} f(x)\sin mx\,dx = \sum_{n=1}^{\infty} a_n\int_a^{a+2\pi}\cos nx\sin mx\,dx + \sum_{n=1}^{\infty} b_n\int_a^{a+2\pi}\sin nx\sin mx\,dx \tag{72}$$

Here the use of the orthogonality conditions 45 and 46 gives

$$b_m = \frac{1}{\pi}\int_a^{a+2\pi} f(x)\sin mx\,dx \tag{73}$$

which is a formula for the coefficients of the sine series in Eq. 66 or 67.

A convenient form for the formulas 70 and 71 is obtained if the fundamental range is specified as $-\pi \leqq x \leqq \pi$, as can always be accomplished by a suitable definition for the independent variable. Then, n being written in place of m to provide a more evident consistency with the form of the series as stated by Eq. 67, the formulas read*

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad \text{for } n = 0, 1, 2, \cdots \tag{74}$$

and

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \sin nx \, dx, \quad \text{for } n = 1, 2, 3, \cdots \tag{75}$$

It is now significant to observe that if the specified function $f(x)$ is *even*, that is, if

$$f(-x) = f(x) \tag{76}$$

the integrand in the integral for b_n is an odd function of x, and the integral between the symmetrical limits $-\pi$ and π vanishes. In other words, all the coefficients b_n are then zero, so that the function $f(x)$ is represented by a cosine series alone. This should, of course, be expected for an even function $f(x)$ since the cosine functions are even and the sine functions odd.

On the other hand if $f(x)$, with regard to the fundamental range, $-\pi \leqq x \leqq \pi$, is an odd function of x, that is, if

$$f(-x) = -f(x) \tag{77}$$

the integrand in the integral for a_n is an odd function, and this integral vanishes for the limits $-\pi$ to π. All the coefficients of the cosine terms are then zero, and the function $f(x)$ is represented by a sine series alone.

This state of affairs suggests that in general one may write

$$f(x) = f_1(x) + f_2(x) \tag{78}$$

in which $f_1(x)$ is an even function and $f_2(x)$ an odd one. That such a representation for the function $f(x)$ should always be possible is clear from the fact that if x in Eq. 78 is replaced by $-x$, and the relations 76 and 77 are observed with regard to the even and odd functions $f_1(x)$ and $f_2(x)$ respectively, there results

$$f(-x) = f_1(x) - f_2(x) \tag{79}$$

*It may not be superfluous to point out that the derivation of these formulas for the Fourier coefficients in no way proves that the series can actually be used to represent a periodic function. The conditions under which such a representation is possible and a discussion of the convergence of the series are given in Arts. 7, 8, 12, 13, 14.

Addition and subtraction of Eqs. 78 and 79 then yield

$$f_1(x) = \tfrac{1}{2}\{f(x) + f(-x)\} \tag{80}$$

and

$$f_2(x) = \tfrac{1}{2}\{f(x) - f(-x)\} \tag{81}$$

Hence the even and odd components $f_1(x)$ and $f_2(x)$ of an arbitrary function $f(x)$ can always be uniquely determined either analytically or graphically.

Substituting the representation 78 for $f(x)$ into the integrals of Eqs. 74 and 75, and taking note of the odd and even character of the functions $f_2(x)$ and $f_1(x)$ respectively, one finds that

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f_1(x) \cos nx\, dx, \quad \text{for } n = 0, 1, 2, \cdots \tag{82}$$

and

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f_2(x) \sin nx\, dx, \quad \text{for } n = 1, 2, 3, \cdots \tag{83}$$

In other words, the coefficients of the cosine terms in the series 66 are determined from the even component of $f(x)$ alone, and the coefficients of the sine terms are determined from the odd component alone. Inasmuch as an arbitrary function $f(x)$ must in general contain both even and odd components according to the decomposition expressed by Eq. 78, it follows that the trigonometric series representing $f(x)$ must contain both sine and cosine terms. When it is assumed, then, that the even and odd functions respectively are completely represented by cosine and sine series, it follows that the form of the assumed series representation for $f(x)$ as given by Eq. 66 is also sufficient. This question, as well as the manner in which the trigonometric series approximates the function $f(x)$ as the number of terms in its partial sum is increased, is discussed in the subsequent articles.

In the meantime it may be useful to observe that the integrands in both the integrals 82 and 83 are even functions of x. Consequently the same value is obtained if the integration is extended only over the range zero to π and the result is multiplied by two. This gives the alternative formulas

$$a_n = \frac{2}{\pi}\int_{0}^{\pi} f_1(x) \cos nx\, dx, \quad \text{for } n = 0, 1, 2, \cdots \tag{84}$$

and

$$b_n = \frac{2}{\pi}\int_{0}^{\pi} f_2(x) \sin nx\, dx, \quad \text{for } n = 1, 2, 3, \cdots \tag{85}$$

In the discussion so far, it has been assumed that the function $f(x)$ is defined only within the interval $a \leqq x \leqq a + 2\pi$, that is, that the function does not necessarily exist also for values of x not in the stated range. For example, if $f(x)$ represents the initial distribution of voltage on a transmission line and the interval $a \leqq x \leqq a + 2\pi$ corresponds, by reason of a suitable change of variable, to the length of that line, it is manifestly clear that an inquiry regarding the values of this function beyond the limits of its fundamental range has no sense. From a practical point of view such an inquiry does not arise since one is content with a solution valid over the extent of the physical system.

It is, on the other hand, sensible to inquire, for the moment out of pure curiosity perhaps, what kind of function the trigonometric series represents when all restrictions on the independent variable are removed, in other words, if x is allowed to take on all values from $-\infty$ to ∞. In view of the periodic character of the trigonometric functions, the answer to this question is obvious. The trigonometric series represents a periodic function with the period 2π, such that the behavior of the function throughout any one period matches that of the function $f(x)$ over its fundamental range.

It is thus clear that the trigonometric series in Eq. 66 is capable of representing an arbitrary *periodic* function over the entire range of its independent variable from minus to plus infinity. Such functions occur quite frequently in engineering analysis, the independent variable usually being the time t. The periodicity of the function is expressed by the relation

$$f(t + k\tau) = f(t) \tag{86}$$

in which τ is the *period* (sometimes referred to as the *fundamental* period) and k is any positive or negative integer. The reciprocal of τ is the fundamental *frequency*

$$f = \frac{1}{\tau} \tag{87}$$

and

$$\omega = 2\pi f = \frac{2\pi}{\tau} \tag{88}$$

is the fundamental *angular* frequency or *radian* frequency.

The Fourier series representation for the periodic function $f(t)$ is written

$$\begin{aligned} f(t) = \frac{a_0}{2} &+ a_1 \cos \omega t + a_2 \cos 2\omega t + \cdots \\ &+ b_1 \sin \omega t + b_2 \sin 2\omega t + \cdots \end{aligned} \tag{89}$$

in which, according to Eqs. 74 and 75,

$$a_n = \frac{\omega}{\pi}\int_{-\pi/\omega}^{\pi/\omega} f(t) \cos n\omega t \, dt, \quad \text{for } n = 0, 1, 2, \cdots \qquad [90]$$

and

$$b_n = \frac{\omega}{\pi}\int_{-\pi/\omega}^{\pi/\omega} f(t) \sin n\omega t \, dt, \quad \text{for } n = 1, 2, 3, \cdots \qquad [91]$$

4. The phase angles of the harmonic components

In view of the trigonometric identities

$$c_n \cos (n\omega t + \phi_n) \equiv c_n \sin \left(n\omega t + \frac{\pi}{2} + \phi_n\right)$$
$$\equiv c_n \cos \phi_n \cos n\omega t - c_n \sin \phi_n \sin n\omega t \qquad [92]$$

it is possible to combine the sine and cosine series in Eqs. 66 or 89 through letting

$$\begin{aligned} a_n &= c_n \cos \phi_n \\ b_n &= -c_n \sin \phi_n \end{aligned} \qquad [93]$$

whence

$$\begin{aligned} c_n &= \sqrt{a_n^2 + b_n^2} \\ \phi_n &= -\tan^{-1}\left(\frac{b_n}{a_n}\right) \end{aligned} \qquad [94]$$

The Fourier series can then be written as a sum of sine or cosine terms alone, for example, as

$$f(t) = \frac{c_0}{2} + c_1 \cos (\omega t + \phi_1) + c_2 \cos (2\omega t + \phi_2) + \cdots \qquad [95]$$

The term having the fundamental angular frequency is commonly called the *fundamental component* of the periodic function $f(t)$, and the remaining terms, whose frequencies are integer multiples of the fundamental, are referred to as *harmonics*. The coefficients c_n and the angles ϕ_n, which are determined by Eqs. 94 together with Eqs. 90 and 91, are known as the *harmonic amplitudes* and *phase angles*.

Because of the periodic nature of the function $f(t)$ it is usually possible to select the origin for the independent variable t at any convenient point. In subsequent manipulations it may, nevertheless, be desirable to shift the origin to a new location. Thus, if the variable t in $f(t)$ is replaced by $(t - t_0)$, the effect is to shift the origin back by an amount equal to t_0 (this shift amounts to retarding the sequence of values of the function

by t_0). Graphically, this change of variable corresponds to shifting the plot of the function *forward* by an amount t_0. A typical harmonic component becomes

$$c_n \cos [n\omega(t - t_0) + \phi_n] = c_n \cos [n\omega t + (\phi_n - n\omega t_0)] \qquad [96]$$

This result may evidently be interpreted as a change in the harmonic phase angle to the new value

$$\phi'_n = \phi_n - n\omega t_0 = \phi_n - \frac{2\pi n t_0}{\tau} \qquad [97]$$

It is significant to note here that each harmonic phase angle is changed by an increment proportional to the order n of the corresponding harmonic component. Such a proportionate change in the harmonic phase angles, therefore, leaves the resultant form of the function $f(t)$ unchanged except for a translation of the function as a whole.

5. Even and odd harmonics

It frequently occurs that the given perodic function satisfies the condition

$$f\left(t \pm \frac{\tau}{2}\right) = -f(t) \qquad [98]$$

which means that the sequence of values of the function throughout any half period are the negatives of the values encountered throughout the preceding or succeeding half period. In order to observe the effect of this condition upon the form of the resulting Fourier series, one may write the expressions 90 and 91 for the coefficients in the form

$$a_n = \frac{\omega}{\pi}\int_0^{\pi/\omega}\left[f(t)\cos n\omega t + f\left(t - \frac{\tau}{2}\right)\cos n\omega\left(t - \frac{\tau}{2}\right)\right]dt \qquad [99]$$

and

$$b_n = \frac{\omega}{\pi}\int_0^{\pi/\omega}\left[f(t)\sin n\omega t + f\left(t - \frac{\tau}{2}\right)\sin n\omega\left(t - \frac{\tau}{2}\right)\right]dt \qquad [100]$$

These are readily recognized as being equivalent to Eqs. 90 and 91 through noting, for example, that

$$\int_{-\pi/\omega}^{0} f(t)\cos n\omega t\, dt = \int_0^{\pi/\omega} f\left(t - \frac{\tau}{2}\right)\cos n\omega\left(t - \frac{\tau}{2}\right)dt \qquad [101]$$

because the second integral is obtained from the first when one makes the substitution $t \to t - \tau/2$ and appropriately changes the limits of integra-

tion. Then, observing that $n\omega\tau/2 = n\pi$, and hence that

$$\begin{aligned} \cos n\omega\left(t - \frac{\tau}{2}\right) &= \cos n\pi \cos n\omega t \\ \sin n\omega\left(t - \frac{\tau}{2}\right) &= \cos n\pi \sin n\omega t \end{aligned} \qquad [102]$$

one finds that the condition 98 substituted into Eqs. 99 and 100 is seen to yield

$$a_n = \frac{\omega}{\pi}(1 - \cos n\pi) \int_0^{\pi/\omega} f(t) \cos n\omega t \, dt \qquad [103]$$

and

$$b_n = \frac{\omega}{\pi}(1 - \cos n\pi) \int_0^{\pi/\omega} f(t) \sin n\omega t \, dt \qquad [104]$$

The factor $(1 - \cos n\pi)$ is zero for all even integers of n and equal to 2 for all odd integer values of n. Hence it is clear that when the periodic function has the property expressed by Eq. 98, its Fourier series contains odd harmonics only. The average value or constant component $a_0/2$ is evidently zero also. The coefficients of the odd harmonics are then given by

$$a_n = \frac{2\omega}{\pi} \int_0^{\pi/\omega} f(t) \cos n\omega t \, dt \qquad [105]$$

and

$$b_n = \frac{2\omega}{\pi} \int_0^{\pi/\omega} f(t) \sin n\omega t \, dt \qquad [106]$$

These considerations may suggest the question of what the results are when the function $f(t)$ satisfies a relation complementary to that expressed by Eq. 98, namely, when

$$f\left(t \pm \frac{\tau}{2}\right) = +f(t) \qquad [107]$$

Although recognizing that the factor $(1 - \cos n\pi)$ in Eqs. 103 and 104 is changed to $(1 + \cos n\pi)$ gives the answer a moment's reflection reveals that the condition 107 merely states that $f(t)$ has the period $\tau/2$ instead of τ. If τ is retained as the fundamental period, the result is that the Fourier series contains only even harmonics, which is just another way of saying that the period of $f(t)$ is actually half as large.

On the basis of these thoughts it appears that any periodic function may be assumed to consist of two components, of which one satisfies the

condition 98 and the other has twice the fundamental frequency. This decomposition may be indicated by

$$f(t) = f_{\mathrm{I}}(t) + f_{\mathrm{II}}(t) \tag{108}$$

in which

$$f_{\mathrm{I}}\left(t \pm \frac{T}{2}\right) = f_{\mathrm{I}}(t) \tag{109}$$

and

$$f_{\mathrm{II}}\left(t \pm \frac{T}{2}\right) = -f_{\mathrm{II}}(t) \tag{110}$$

It then follows that

$$f\left(t \pm \frac{T}{2}\right) = f_{\mathrm{I}}(t) - f_{\mathrm{II}}(t) \tag{111}$$

and hence, adding and subtracting Eqs. 108 and 111, that

$$f_{\mathrm{I}}(t) = \frac{1}{2}\left\{f(t) + f\left(t \pm \frac{T}{2}\right)\right\} \tag{112}$$

$$f_{\mathrm{II}}(t) = \frac{1}{2}\left\{f(t) - f\left(t \pm \frac{T}{2}\right)\right\} \tag{113}$$

This decomposition is similar in analytic form to the decomposition into even and odd components as stated by Eqs. 80 and 81, but should not be confused with the latter. The decomposition into even and odd components according to Eqs. 80 and 81 is a decomposition of the given function into components which are respectively symmetrical and antisymmetrical about the ordinates at $x = 0$ or $t = 0$, and hence yields a decomposition of the Fourier series into its component cosine and sine series, each of which in general contains both even and odd harmonics, and hence can be further decomposed according to the relations 112 and 113.

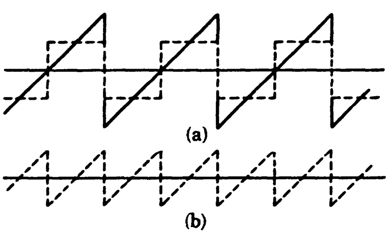


FIG. 3. Decomposition of a saw-tooth wave into even and odd harmonics

The component functions given by Eqs. 112 and 113, on the other hand, may individually be either even or odd but in general are neither, and hence can also be further decomposed. The decomposition indicated by Eq. 108 is such that the first component contributes the even har-

monic components to the resulting Fourier series, and the second contributes only the odd harmonics. An interesting example of this decomposition is illustrated for the saw-tooth wave shown in part (a) of Fig. 3. Here the rectangular wave (shown dotted) represents the odd harmonic components, and the even harmonics are due to the saw-tooth wave of double frequency, part (b), which remains after the rectangular component is subtracted from the original saw-tooth wave.

6. Alternative Fourier expansions for a function defined over a finite range

The preceding articles show that there is an infinite variety of ways to establish a Fourier series representation for a function which is defined over a finite region, if the sole object is to obtain a trigonometric series which yields the correct values of the stated function over this finite range only. One way is to consider the defining range of the given function as the fundamental period, as is done in Art. 3. However, since this range may alternatively be considered as only a part of the fundamental period, and since the definition of the given function over the remainder of this period as well as the extent of the period are entirely arbitrary, it is clear that any number of trigonometric series representations may be found. All of them yield the same values over that portion of the period which corresponds to the original defining range, although they may show a variety of behaviors throughout the remainder of each period.

It should be observed that if the determination of a Fourier expansion appears as part of the process of solving a boundary value problem, such a variety of alternative possible procedures does not exist, because the conditions of the problem permit the choice of only one set of proper functions in terms of which the expansion may be made. There are, however, many other ways in which Fourier expansions enter into engineering analysis, and in most of them the object of using the series representation is solely to get an approximating function in the form of a trigonometric series for some given function whose behavior is specified over a finite interval. Inasmuch as the engineer must, for practical reasons, limit the expansion to a finite number of terms, and the computational labor as well as other economic factors dictates that this finite number shall be as small as possible, the freedom of choice mentioned in the preceding paragraph must be recognized as an important consideration in the selection of a suitable process of analysis.

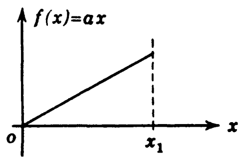


Fig. 4. A function defined over a limited range.

As a simple illustration of how a variety of series representations may

be obtained for a given function defined over a finite range, it is interesting to consider the function illustrated in Fig. 4 for which the defining range is the region $0 \leqq x \leqq x_1$. Parts (a) to (e) of Fig. 5 show several

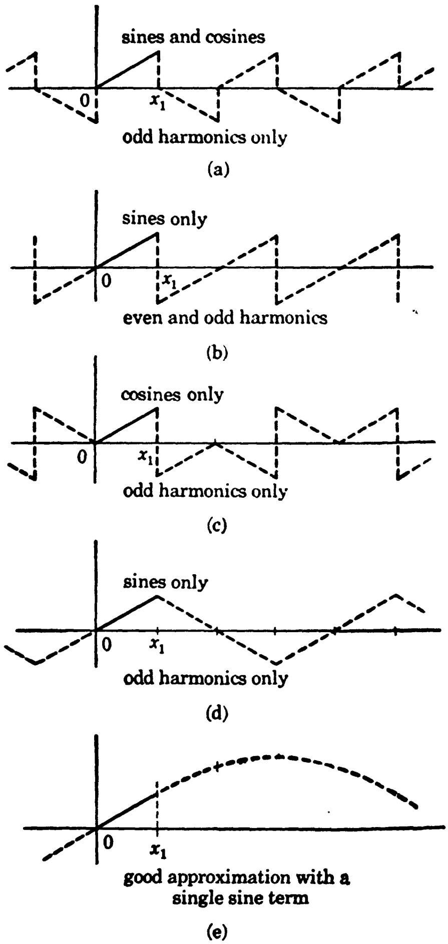


FIG. 5. Various possible periodic continuations of the function of Fig. 4.

ways in which this function may be assumed to be continued beyond the defining range so as to yield a periodic function. In each case the behavior over the defining range is the same as that given in Fig. 4, but the series representations for the individual cases are nevertheless quite different, as indicated on each sketch.

It is also significant that the rate of convergence of the resulting Fourier series may be quite different for the different forms of periodic functions. As is also pointed out in a subsequent article, the convergence is in general more rapid for a smooth function than it is for one which varies erratically or has large first or higher derivatives. For example, the series for a function which has discontinuities, like those sketched in parts (a) to (c) of Fig. 5, in general converges rather slowly, so that a large number of terms must be calculated in order to establish a fairly good representation for the function. If the latter is continuous, as it is in parts (d) and (e) of the figure, the convergence is considerably more rapid. The function shown in part (e) has a more rapidly converging series than the function in part (d) whose first derivative has discontinuities. In fact, the function in part (e) is approximated very well by a single sine term.

7. The Fourier series as a special form of the Laurent expansion; the complex Fourier series

The Fourier series may alternatively be obtained from the Laurent expansion discussed in Art. 12 of Ch. VI. According to Eqs. 163 and 166 of that chapter, the expansion has the form

$$f(z) = \sum_{n=-\infty}^{\infty} \alpha_n (z - z_0)^n \qquad [114]$$

in which the coefficients are given by the formula

$$\alpha_n = \frac{1}{2\pi j} \oint_S \frac{f(\zeta)\, d\zeta}{(\zeta - z_0)^{n+1}} \qquad [115]$$

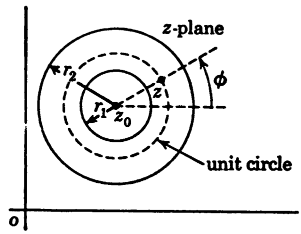


Fig. 6. Region of uniform convergence of a Laurent series from which the Fourier series may be obtained by an appropriate change of variable.

The region of uniform convergence lies between the circles r_1 and r_2 indicated in Fig. 6, for which it is assumed that $r_1 < 1 < r_2$. The expansion 114 then converges uniformly for all points on the unit circle drawn with z_0 as a center. As discussed in the previous chapter, this statement assumes, of course, that the function is regular and continuous at all points within the annular region enclosed by the circles with radii r_1 and r_2. No restriction is implied by the assumption that $r_1 < 1 < r_2$, since this condition may in any case be obtained by means of an obvious change of variable, should it not be met in the first place.

If the point z in the expansion 114 is assumed to refer to any point on the unit circle, and the latter is also chosen as the path of integration S in the formula 115, one may write

$$z - z_0 = e^{j\phi} \tag{116}$$

and

$$\zeta - z_0 = e^{j\psi} \tag{117}$$

whence

$$\frac{d\zeta}{\zeta - z_0} = j\, d\psi \tag{118}$$

The expansion 114 then takes the form

$$f(z) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jn\phi} = f(\phi) \tag{119}$$

in which, according to Eqs. 115, 117, and 118,

$$\alpha_n = \frac{1}{2\pi j}\int_a^{a+2\pi} \frac{f(\psi) j\, d\psi}{e^{jn\psi}} = \frac{1}{2\pi}\int_a^{a+2\pi} f(\psi)e^{-jn\psi}\, d\psi \tag{120}$$

Since the point z is restricted to lie on a circle about z_0, $f(z)$ evidently reduces to a function of the real variable ϕ. In order to conform with more familiar conventions, the symbol ϕ may be replaced by x and the variable of integration ψ by ξ, so that Eqs. 119 and 120 take the form

$$f(x) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jnx} \tag{121}$$

and

$$\alpha_n = \frac{1}{2\pi}\int_a^{a+2\pi} f(\xi)e^{-jn\xi}\, d\xi \tag{122}$$

Although the function $f(x)$ is real, it should be observed that the coefficients α_n are complex. The series representation for the function $f(x)$, as expressed by Eq. 121, is known as the *complex Fourier series.* It may readily be shown to be entirely equivalent to the real form in terms of sines and cosines. The first step in this regard is the observation that, according to Eq. 122, the coefficient α_n is replaced by its conjugate value if n is replaced by $-n$, that is,

$$\alpha_{-n} = \bar{\alpha}_n \tag{123}$$

This result is clear from the fact that changing the sign of n in Eq. 158 is equivalent to changing the sign of j. The series 121 may be written

out in the form

$$f(x) = \alpha_0 + \begin{cases} \alpha_1 e^{jx} + \alpha_2 e^{j2x} + \alpha_3 e^{j3x} + \cdots \\ \alpha_{-1} e^{-jx} + \alpha_{-2} e^{-j2x} + \alpha_{-3} e^{-j3x} + \cdots \end{cases} \qquad [124]$$

from which it is seen that the terms appear in pairs of conjugates. A typical conjugate pair reads

$$\alpha_n e^{jnx} + \alpha_{-n} e^{-jnx} \qquad [125]$$

Now if one writes

$$\alpha_n = \frac{a_n - jb_n}{2} \qquad [126]$$

then, according to Eq. 123,

$$\alpha_{-n} = \frac{a_n + jb_n}{2} \qquad [127]$$

The pair of terms given in Eq. 125 then becomes

$$\alpha_n e^{jnx} + \alpha_{-n} e^{-jnx} = a_n \cos nx + b_n \sin nx \qquad [128]$$

whereas the separation of Eq. 122 into its real and imaginary parts yields

$$a_n = \frac{1}{\pi} \int_a^{a+2\pi} f(\xi) \cos n\xi \, d\xi \qquad [129]$$

and

$$b_n = \frac{1}{\pi} \int_a^{a+2\pi} f(\xi) \sin n\xi \, d\xi \qquad [130]$$

Since $b_0 = 0$, and hence $\alpha_0 = \frac{a_0}{2}$, Eq. 128 shows that the series 124 is equivalent to

$$f(x) = \frac{a_0}{2} + a_1 \cos x + a_2 \cos 2x + \cdots$$
$$+ b_1 \sin x + b_2 \sin 2x + \cdots \qquad [131]$$

which is the familiar form for the Fourier series in terms of its sine and cosine components. The formulas 129 and 130 are seen to be identical with those given by Eqs. 70 and 73. The equivalence between the complex form 121 and the ordinary form for the Fourier series is thus established, and the formula 115 for the coefficients of the Laurent expansion is seen to contain the formulas for the Fourier coefficients as a special case.

The complex form for the Fourier series is more convenient than the ordinary form for many manipulations because of its relative compactness as contrasted with the sine and cosine form, and also because of the

greater ease with which the exponential function may be manipulated with regard to various algebraic as well as differential and integral operations. In this connection it may be observed, incidentally, that the complex form requires only the one formula 122 for the evaluation of its coefficients as contrasted to the two formulas, 129 and 130, which are ordinarily needed.

The relation between the complex coefficient α_n and the real coefficients a_n and b_n is further illustrated by the sketch in Fig. 7 which shows a pair of conjugate complex coefficients α_n and α_{-n}. These may be regarded as the conjugate complex terms $\alpha_n e^{jnx}$ and $\alpha_{-n}e^{-jnx}$ for $x = 0$. As the variable x increases from the value zero, the vectors which these terms represent rotate in opposite directions at an angular rate of n radians per unit of x. They, therefore, remain conjugates for all values of x, so that their vector sum becomes a real, simple harmonic function of x. For $x = \omega t$ this pair of conjugate terms is thus seen to represent a simple harmonic function of the time t with an angular frequency of $n\omega$ radians per second.

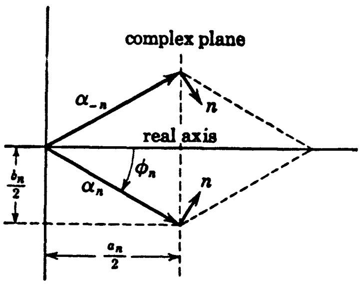


FIG. 7. A pair of conjugate complex coefficients of the complex Fourier series.

Denoting the angle of the complex coefficient α_n by ϕ_n, as indicated in Fig. 7, that is, writing

$$\alpha_n = |\alpha_n|\, e^{j\phi_n}, \quad \alpha_{-n} = |\alpha_n|\, e^{-j\phi_n} \tag{132}$$

one finds that the expression for this simple harmonic function is given by

$$\begin{aligned} \alpha_n e^{jnx} + \alpha_{-n}e^{-jnx} &= |\alpha_n|(e^{j(nx+\phi_n)} + e^{-j(nx+\phi_n)}) \\ &= 2|\alpha_n| \cos (nx + \phi_n) \end{aligned} \tag{133}$$

Equations 126 and 132 give

$$a_n - jb_n = 2|\alpha_n| \cos \phi_n + j2|\alpha_n| \sin \phi_n \tag{134}$$

or separating reals and imaginaries this yields

$$\begin{aligned} a_n &= 2|\alpha_n| \cos \phi_n \\ b_n &= -2|\alpha_n| \sin \phi_n \end{aligned} \tag{135}$$

Comparison with Eqs. 93 and 95, shows that the resulting harmonic amplitude c_n is equal to *twice* the magnitude of the complex Fourier coefficient α_n, whereas the harmonic phase angle is the angle ϕ_n of the

complex coefficient α_n. These results are also evident from inspection of Fig. 7.

The complex form for the Fourier series is thus seen to contain the phase angles of its various harmonic components by virtue of the complex character of the complex coefficients. The formula 122 for these coefficients yields these harmonic phase angles as well as the corresponding amplitudes $2|\alpha_n|$, and the complex form of the series, as expressed by Eq. 121, contains both the amplitudes and the phase angles although the latter do not appear explicitly. This fact, namely that the harmonic phase angles are *implicitly* contained in the complex form 121, and hence need not be explicitly written down during any manipulations which may be carried out with the series, is one of its greatest labor-saving virtues when the Fourier series is used in various analytic formulations. This advantage becomes particularly apparent from the discussions given in Art. 9.

Inasmuch as the basis for the Laurent expansion requires that the function $f(z)$ be regular and continuous throughout the annular region enclosed by the circle with radii r_1 and r_2 in Fig. 6, one should expect that the validity of the Fourier series would be similarly restricted. Although this restriction, in general, holds if one expects uniform convergence of the series for all values of its independent variable, it is found nevertheless that the conditions for the possibility of obtaining a Fourier series representation for a given function are somewhat less confining if certain reservations are accepted regarding the convergence of the resulting series.

The conditions, known as the *Dirichlet conditions*, under which a Fourier series representation for a given function is possible, state that throughout the fundamental range for which the function is defined, it shall possess a finite number of maxima and minima, and in spite of having a finite number of discontinuities and points where the function becomes infinite, it shall possess an absolutely convergent integral; that is,*

$$\int_a^b |f(x)|\, dx \qquad [136]$$

shall be finite. If, according to the latter part of this statement, the function becomes infinite at some point, this infinity shall be integrable, for example, like a logarithmic infinity. (It should be recalled that the area

*The absolute convergence is required only if the integral $\int_a^b f(x)\, dx$ becomes improper. Actually, the conditions set down by Dirichlet are more stringent than those just stated, and do not permit the function $f(x)$ to become unbounded. The conditions stated here are, nevertheless, designated by many writers as the Dirichlet conditions, and inasmuch as it is a convenient way of referring to them, this somewhat inaccurate practice is used here also.

under a logarithmic infinity is finite.) The series converges uniformly for all values of the independent variable except those corresponding to points of discontinuity, and at the points corresponding to infinite values for the function the series can, of course, no longer converge.

At a point of discontinuity, which may be denoted by $x = x_1$, the Fourier series yields the average of the two values of the function immediately adjacent to this discontinuity, that is, the series yields the value

$$\frac{f(x_1 - 0) + f(x_1 + 0)}{2} \qquad [137]$$

The reason for this property of the series is discussed in Arts. 12 and 13. Together with a number of other interesting items, it is illustrated by some of the examples of the next article.

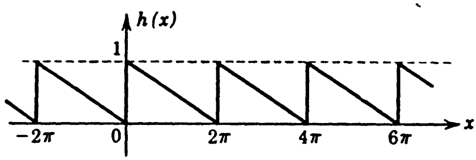


FIG. 8. The saw-tooth wave is basic in the description of discontinuities.

8. Several illustrative examples; a criterion regarding the rate of convergence

The first example to be considered is shown in Fig. 8. This function $h(x)$ has discontinuities of unit magnitude which occur at the origin and at integer multiples of 2π. It is similar to the saw-tooth wave shown in Fig. 3 except that it is turned over and raised so as to lie upon the x-axis. This function may be defined by the statements

$$\begin{aligned} h(x) &= \frac{1}{2} - \frac{\pi + x}{2\pi} \quad \text{for } -\pi < x < 0 \\ h(x) &= \frac{1}{2} + \frac{\pi - x}{2\pi} \quad \text{for } 0 < x < \pi \end{aligned} \qquad [138]$$

It is observed, therefore, that except for the constant component ½, the function is odd. Hence its Fourier series is given by ½ plus a sine series whose coefficients are evaluated by the formula 85. These coefficients are, therefore, given by

$$b_n = \frac{1}{\pi^2} \int_0^{\pi} (\pi - x) \sin nx \, dx \qquad [139]$$

The integration yields

$$b_n = \frac{1}{n\pi} \tag{140}$$

so that the resulting Fourier series becomes

$$h(x) = \frac{1}{2} + \frac{1}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \cdots\right] \tag{141}$$

At the points $x = 0, 2\pi, 4\pi, \cdots -2\pi, -4\pi, \cdots$ etc., the value of the series is $\frac{1}{2}$, which is the arithmetic mean between the values of $h(x)$ immediately adjacent to these points of discontinuity as stated by the expression 137. For the immediate vicinity of these points the series cannot be said to converge uniformly because it yields different values, depending upon the direction from which these points are approached, and hence the partial sums do not converge toward a definite limit in this vicinity as n is indefinitely increased.

In order to investigate the convergence for other values of x one may regard the partial sum

$$s_n(x) = \frac{1}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 2x}{2} + \cdots + \frac{\sin nx}{n}\right] \tag{142}$$

and form

$$\begin{aligned}&|s_{n+k}(x) - s_n(x)|\\ &\quad = \frac{1}{\pi}\left|\frac{\sin (n+1)x}{n+1} + \frac{\sin (n+2)x}{n+2} + \cdots + \frac{\sin (n+k)x}{n+k}\right|\end{aligned} \tag{143}$$

Letting

$$\sigma_n = \sin x + \sin 2x + \cdots + \sin nx \tag{144}$$

one may write the expression 143

$$\begin{aligned}&|s_{n+k}(x) - s_n(x)|\\ &\quad = \frac{1}{\pi}\left|\frac{\sigma_{n+1} - \sigma_n}{n+1} + \frac{\sigma_{n+2} - \sigma_{n+1}}{n+2} + \cdots + \frac{\sigma_{n+k} - \sigma_{n+k-1}}{n+k}\right|\end{aligned} \tag{145}$$

The right-hand side of this equation may be written alternatively in the form

$$\begin{aligned}\frac{1}{\pi}\Bigg|\frac{-\sigma_n}{n+1} &+ \sigma_{n+1}\left(\frac{1}{n+1} - \frac{1}{n+2}\right) + \sigma_{n+2}\left(\frac{1}{n+2} - \frac{1}{n+3}\right) + \cdots\\ &+ \sigma_{n+k-1}\left(\frac{1}{n+k-1} - \frac{1}{n+k}\right) + \frac{\sigma_{n+k}}{n+k}\Bigg|\end{aligned} \tag{146}$$

Now, according to Eq. 12,

$$\sigma_n = \frac{\cos\frac{x}{2} - \cos(2n+1)\frac{x}{2}}{2\sin\frac{x}{2}} \qquad [147]$$

Although, for various values of n and x, this sum σ_n may have a variety of positive or negative values (in particular for $x = 0$ or a multiple of π it has the value zero), it is clear that a finite positive quantity S may be found which the absolute value of σ_n cannot exceed for any x and any n, no matter how large the value of the latter may be chosen. In other words, it is possible to specify that

$$|\sigma_n| < S \quad \text{for any } x \text{ or } n \text{ values} \qquad [148]$$

in which S is finite.

Observing that the coefficients of the partial sums $\sigma_{n+1}, \sigma_{n+2}, \cdots \sigma_{n+k}$ in the expression 146 are positive, and that σ_n may have a numerically negative value while the remaining partial sums are at the same time all positive, one sees that, even in these most unfavorable circumstances, this expression nevertheless cannot have a value in excess of

$$\frac{S}{\pi}\left|\frac{1}{n+1}+\frac{1}{n+1}-\frac{1}{n+2}+\frac{1}{n+2}-\frac{1}{n+3}+-\cdots-\frac{1}{n+k}+\frac{1}{n+k}\right| = \frac{2S}{\pi(n+1)} \qquad [149]$$

Hence it is established that

$$|s_{n+k}(x) - s_n(x)| < \frac{2S}{\pi(n+1)} \quad \text{for all } n \text{ and } k \qquad [150]$$

Consequently the statement that

$$|s_{n+k}(x) - s_n(x)| < \epsilon \quad \text{independently of } x \qquad [151]$$

for all values of k and all values $n > N$ may be justified through choosing

$$\frac{2S}{\pi(n+1)} < \epsilon \qquad [152]$$

or

$$(n+1) > \frac{2S}{\pi\epsilon} \qquad [153]$$

which means

$$N = \frac{2S}{\pi\epsilon} - 1 \qquad [154]$$

According to Cauchy's principle of convergence (Eq. 94, Art. 9, Ch. VI),

therefore, the Fourier series 141 for the function illustrated in Fig. 8 is seen to converge uniformly over any range of x-values which excludes the points of discontinuity.

The results obtained for this simple example enable one to draw conclusions regarding the convergence of Fourier series for more arbitrary functions possessing a finite number of discontinuities. As a first step toward evolving such a generalization, a function $f(x)$ may be assumed to have a single discontinuity equal to the value δ at some point $x = x_\nu$ in the fundamental range $a \leqq x \leqq a + 2\pi$. In other words, the function $f(x)$ is continuous throughout this range except for a sudden jump δ (positive or negative) in its value at the one point $x = x_\nu$.

In terms of the function $h(x)$ of Fig. 8, one may form a function $\delta \cdot h(x - x_\nu)$ which is a saw-tooth wave having a jump equal to δ at the point $x = x_\nu$ in the fundamental range. If this function is subtracted from $f(x)$, the resulting function

$$F(x) = f(x) - \delta h(x - x_\nu) \tag{155}$$

must be continuous throughout the fundamental range $a \leqq x \leqq a + 2\pi$.

With this result tucked away in one's mind for future reference, attention is for the moment turned toward the formula 122 for the complex Fourier coefficients, which is repeated here for the convenience of the reader

$$\alpha_n = \frac{1}{2\pi}\int_a^{a+2\pi} f(\xi)e^{-jn\xi}\,d\xi \tag{156}$$

Here one may apply the principle of integration by parts, for which the well-known formula reads

$$\int u\,dv = uv - \int v\,dv \tag{157}$$

Letting

$$u = f(\xi) \qquad \text{and} \qquad dv = e^{-jn\xi}\,d\xi \tag{158}$$

one has

$$du = f^{(1)}(\xi)\,d\xi \qquad \text{and} \qquad v = -\frac{e^{-jn\xi}}{jn} \tag{159}$$

in which the superscript (1) on the function $f(\xi)$ indicates that its first derivative is meant. Subsequently, the superscripts (2), (3), etc., are used to denote the second and higher order derivatives.

Substitution into the formula 157 yields for the coefficient α_n of Eq. 156

$$\alpha_n = -\frac{f(\xi)e^{-jn\xi}}{2\pi nj}\Bigg]_a^{a+2\pi} + \frac{1}{2\pi nj}\int_a^{a+2\pi} f^{(1)}(\xi)e^{-jn\xi}\,d\xi \tag{160}$$

However,

$$\left.\frac{f(\xi)e^{-jn\xi}}{2\pi nj}\right]_a^{a+2\pi} = \frac{e^{-jna}}{2\pi nj}\{f(a+2\pi) - f(a)\} = 0 \qquad [161]$$

By repeatedly applying the same process, one may obtain the formula

$$\alpha_n = \frac{1}{2\pi(nj)^k}\int_a^{a+2\pi} f^{(k)}(\xi)e^{-jn\xi}\,d\xi \qquad [162]$$

which, of course, is a proper integral only so long as the kth derivative $f^{(k)}(\xi)$ remains finite, although it may be discontinuous.

It may now be supposed that the given function $f(x)$ and all its successive derivatives $f^{(1)}(x)$, $f^{(2)}(x)$, $\cdots$ $f^{(k-1)}(x)$ are continuous, but that the kth derivative $f^{(k)}(x)$ possesses the discontinuity δ at $x = x_\nu$. Then, according to the argument leading to Eq. 155, the function

$$F^{(k)}(x) = f^{(k)}(x) - \delta \cdot h(x - x_\nu) \qquad [163]$$

is still continuous, and hence for it the procedure leading to the integral 162 can be continued at least one step further. Hence the Fourier coefficients α_n for the function $F^{(k)}(x)$ must decrease in magnitude for large values of n at least as rapidly as the ratio $1/(n)^{k+1}$. Inasmuch as the illustrative example at the beginning of this article shows, however, that the Fourier coefficients for the function $h(x - x_\nu)$ decrease with increasingly large n only as fast as the ratio $1/n$, it follows that those for the function $f^{(k)}(x)$ can do no better although they are certain to do as well. These coefficients are given by

$$\alpha_n^{(k)} = \frac{1}{2\pi}\int_a^{a+2\pi} f^{(k)}(\xi)e^{-jn\xi}\,d\xi \qquad [164]$$

and hence the coefficients for the function $f(x)$ are, according to Eq. 162, expressible as

$$\alpha_n = \frac{\alpha_n^{(k)}}{(jn)^k} \qquad [165]$$

One may, therefore, conclude that for a function $f(x)$ which, together with all its successive derivatives up to that of the kth order, is continuous throughout the fundamental range, and for which, therefore, the kth derivative is the first one possessing a discontinuity within this range, the Fourier coefficients are found to decrease in magnitude for large values of n as rapidly as the ratio $1/n^{k+1}$.

Furthermore, one may conclude that, inasmuch as the Fourier series for $h(x - x_\nu)$ converges uniformly over any region whose boundaries exclude the points $x_\nu \pm$ an integer number of 2π's, the series for the function $f^{(k)}(x)$ also converges uniformly over the same ranges. In the

immediate vicinity of a point of discontinuity this series no longer converges uniformly, but its partial sums $s_n^{(k)}(x)$ in absolute value evidently remain smaller than some finite upper bound like the value S for the partial sum 144 or its closed form 147. Recognizing that the Fourier coefficients for the function $f^{(k-1)}(x)$ are those for the function $f^{(k)}(x)$ multiplied by $1/jn$, and following a procedure identical to that used in the investigation of the convergence of $h(x)$, one is led to the conclusion that the series for $f^{(k-1)}(x)$ converges uniformly even over a range which includes the discontinuity of $f^{(k)}(x)$. The series for the derivatives of lower order, and so on down to that for the function $f(x)$, then certainly do likewise.

An important result of this line of thought is the fact that if a given function $f(x)$ is itself continuous but possesses a discontinuity in its first derivative at some point x_ν, the Fourier series for $f(x)$ still converges uniformly at the point $x = x_\nu$. It may, therefore, unquestionably be integrated term by term, and it may also be differentiated term by term, although the point of discontinuity x_ν must then be excluded from the region of uniform convergence.

Each time the Fourier series is integrated term by term, an additional factor n is introduced into the denominator of the expression for α_n, and each time it is differentiated, a factor of n is canceled out of the denominator of this expression. It is readily seen, therefore, that integration improves the rate of convergence of the Fourier series, whereas differentiation makes the series converge more slowly. This fact may have been expected in the first place since integrating a function makes it smoother whereas differentiating it accentuates any existing irregularities.

It is a simple matter to extend these conclusions to include functions having a larger number of discontinuities, as long as this number remains finite. For example, if the function $f(x)$ has the discontinuities $\delta_1, \delta_2, \cdots \delta_\nu$ at the points $x_1, x_2, \cdots x_\nu$ within the fundamental range, the function

$$F(x) = f(x) - \sum_{i=1}^{\nu} \delta_i h(x - x_i) \qquad [166]$$

is continuous throughout this range, whereas the Fourier series for that part of $f(x)$ represented by

$$\sum_{i=1}^{\nu} \delta_i h(x - x_i) \qquad [167]$$

is given by the sum of ν series, each of which has the same general form as that for $h(x)$. The Fourier series for $f(x)$, therefore, converges uniformly over ranges whose boundaries exclude the finite number of discontinuities, and the coefficients decrease in magnitude with large values of n as fast

as and no faster than the ratio $1/n$. The series may be integrated term by term, yielding a series which converges uniformly throughout the fundamental range, but it cannot be differentiated term by term since the resulting series then no longer converges at all.

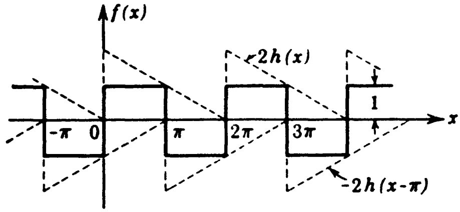


FIG. 9. Two saw-tooth waves add up to yield a square wave.

The second example to be considered is the rectangular wave with identical positive and negative half cycles of unit magnitude as shown in Fig. 9. This function has a discontinuity equal to 2 at the beginning of its fundamental range $0 \leqq x \leqq 2\pi$, and another equal to -2 at the center of this range. The saw-tooth functions representing these two discontinuities are, therefore, given by

$$2h(x) \qquad \text{and} \qquad -2h(x-\pi) \tag{168}$$

After subtracting these from the given function in this example, one finds that there is nothing left, that is,

$$f(x) - 2h(x) + 2h(x-\pi) = 0 \tag{169}$$

Hence, according to Eq. 141, the Fourier series is given by

$$\begin{aligned} f(x) &= \frac{2}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 2x}{2} + \cdots\right] \\ &\quad - \frac{2}{\pi}\left[\frac{\sin (x-\pi)}{1} + \frac{\sin 2(x-\pi)}{2} + \cdots\right] \end{aligned} \tag{170}$$

With

$$\sin n(x-\pi) = (-1)^n \sin nx \tag{171}$$

this result yields for the Fourier series of the square wave shown in Fig. 9

$$f(x) = \frac{4}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots\right] \tag{172}$$

As one should expect from the fact that $f(x)$ is odd, the representation is given by a sine series alone, and since the function satisfies the condi-

tion 98, only the odd harmonics are present. When the origin is shifted to the point $x = \pi/2$, the function becomes an even one and the series is converted into a cosine series. This shift is accomplished by replacing x in Eq. 172 by $x + \pi/2$, thus

$$f\left(x + \frac{\pi}{2}\right) = \frac{4}{\pi}\left[\frac{\cos x}{1} - \frac{\cos 3x}{3} + \frac{\cos 5x}{5} - \cdots\right] \qquad [173]$$

in which it is to be observed that the algebraic signs alternate. It may also be noted from Eq. 172 that, at the points of discontinuity of the function, the series yields the value zero which again is the algebraic mean between the immediately adjacent values of $f(x)$.

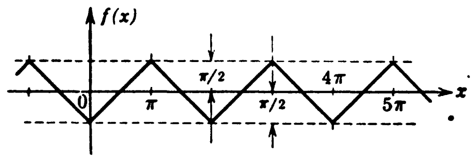


FIG. 10. A triangular wave which results when the rectangular wave of Fig. 9 is integrated.

Integrating the series 172 term by term, one obtains the Fourier series representation for the triangular wave shown in Fig. 10 which is recognized as the result of integrating the square wave of Fig. 9. Thus, for the function of Fig. 10,

$$f(x) = -\frac{4}{\pi}\left[\frac{\cos x}{1} + \frac{\cos 3x}{9} + \frac{\cos 5x}{25} + \cdots\right] \qquad [174]$$

A point which may be a bit puzzling here is the fact that no constant term appears in the expression 174, whereas if one mentally visualizes the function which represents the area under the square wave of Fig. 9, starting the integration at $x = 0$, this function evidently turns out to be a triangular wave lying upon the x-axis, that is, having the form of the wave shown in Fig. 10 but with an additive constant component equal to $\pi/2$. The reason that this constant component is missing from the representation 174 is, of course, the fact that this result is the *indefinite* integral of the series 172 and not the integral from zero to some variable point x. Incidentally, the definite integral from $\pi/2$ to any variable point x has no constant component, as is readily recognized from inspection of the Fig. 9. At all events one must recognize that the indefinite integral can yield only what is customarily referred to as the *alternating component* of the resulting function, and this in the present example is the triangular wave of Fig. 10.

It should be particularly noted that the series 174 converges considerably more rapidly than the series 172 or 173. Thus the coefficients in the series 174 vary as $1/n^2$ whereas those in the series 172 vary only as $1/n$. This observation agrees with what should be expected from the preceding discussion inasmuch as the triangular-wave function is continuous whereas its first derivative, the square-wave function, possesses discontinuities.

9. The Fourier spectrum

The interpretations given in the present article are most commonly used in connection with functions which are periodic with respect to the time t. The essential manipulations are, moreover, most effectively carried out in terms of the complex form of the Fourier series, which is expressed by Eq. 121 together with the formula 122. For $x = \omega t$ these expressions are usually written

$$f(t) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jn\omega t} \qquad [175]$$

with

$$\alpha_n = \frac{\omega}{2\pi} \int_{-\pi/\omega}^{\pi/\omega} f(t) e^{-jn\omega t}\, dt \qquad [176]$$

Here it is useful to regard the general coefficient α_n as a function of the variable $n\omega$, which is the angular frequency of the harmonic component of order n. Since n assumes only integer values, the variable $n\omega$ is a discontinuous one, and the function α_n, which may alternatively be denoted as $\alpha(n\omega)$, has values only for discrete values of its independent variable. Nevertheless, it is useful to regard $\alpha(n\omega)$ as a function rather than merely as denoting the values of the Fourier coefficients, since this view leads to an interesting and useful interpretation of the relations 175 and 176.

The variable t is regarded as the independent variable in a certain region designated as the *time domain*, whereas the variable $n\omega$ represents an independent variable in a corresponding region known as the *frequency domain*. The function which is given as $f(t)$ represents a specification of some desired function in the time domain, and the corresponding function $\alpha(n\omega)$, or α_n, found from the integral Eq. 176, is regarded as the specification of that same function in the frequency domain. In other words, the function $\alpha(n\omega)$ is in every respect just as complete and specific a representation of the desired function as is $f(t)$, the only difference being that $\alpha(n\omega)$ represents that function in a different domain, namely that domain in which $n\omega$ instead of t is the independent variable.

This view that the function $\alpha(n\omega)$ is entirely and uniquely the equivalent of the function $f(t)$ is indisputably tenable since the relation 175 uniquely converts the function $\alpha(n\omega)$ into the function $f(t)$, whereas the relation 176 does the reverse. In other words, these two expressions are a pair of mutually inverse relations in the sense that one of them *undoes* what the other *does*. This circumstance may be placed even more clearly in evidence through substituting the expression 176 for α_n into 175, which operation incidentally requires writing for the variable of integration t in Eq. 176 some other symbol u so as to avoid confusion with the independent variable t in the expression 175, thus

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{\omega}{2\pi} \int_{-\pi/\omega}^{\pi/\omega} f(u)e^{jn\omega(t-u)}\, du \qquad [177]$$

By means of this expression, in which the variables t and u both refer to the time domain, the given function $f(t)$ is expressed in terms of itself.

The pair of relations 175 and 176 may, in the light of this interpretation, be regarded as *transformations* which accompany the change of variable from t to $n\omega$ or vice versa. For this reason the function $\alpha(n\omega)$ is sometimes spoken of as the *Fourier transform* of $f(t)$, and the latter as the *inverse transform* of $\alpha(n\omega)$. In their present form the transformations 175 and 176 are restricted by the condition that the function $f(t)$ must be a periodic function defined over the entire time domain $-\infty < t < \infty$. The removal of this restriction and the accompanying modification in the Fourier transforms are discussed in Art. 19.

Just as the function $f(t)$ may be represented graphically by being plotted in the time domain, so the equivalent function $\alpha(n\omega)$ may be plotted in the frequency domain. The latter domain is more commonly known as the frequency *spectrum* and the plot of $\alpha(n\omega)$ as the *spectrum representation* of the function $f(t)$.

Inasmuch as the function $\alpha(n\omega)$ is in general complex, it is necessary to make two plots in order to represent it completely. The ones usually chosen are the magnitude $|\alpha_n|$ and the angle ϕ_n, although the real and imaginary parts could also be chosen. Plots of $|\alpha_n|$ and ϕ_n are called respectively the *amplitude* and *phase* spectra of the given function $f(t)$. Since $n\omega$ assumes only discrete values, these plots are not in the form of continuous curves but consist merely of a series of vertical lines representing the ordinates of the functions $|\alpha_n|$ and ϕ_n corresponding to integer values of n. For this reason they are referred to as *line spectra*.

The square-wave function of Fig. 9, for example, yields, according to the formula 176,

$$\alpha_n = \frac{\omega}{2\pi}\left\{ -\int_{-\pi/\omega}^{0} e^{-jn\omega t}\, dt + \int_{0}^{\pi/\omega} e^{-jn\omega t}\, dt \right\} \qquad [178]$$

or, the change of variable $t \rightarrow t - \pi/\omega$ being made in the first integral,

$$\alpha_n = \frac{\omega}{2\pi}(1 - e^{jn\pi}) \int_0^{\pi/\omega} e^{-jn\omega t}\, dt \qquad [179]$$

Integrating and substituting limits give

$$\alpha_n = \frac{(1 - e^{jn\pi})(1 - e^{-jn\pi})}{2\pi nj} = \frac{(1 - \cos n\pi)}{\pi nj} \qquad [180]$$

Hence

$$\alpha_n = \begin{cases} 0 & \text{for } n \text{ even} \\ \dfrac{2}{j\pi n} & \text{for } n \text{ odd} \end{cases} \qquad [181]$$

Substituting this result into the sum 175 and combining the conjugate terms yield the series

$$f(t) = \frac{4}{\pi}\left[\frac{\sin \omega t}{1} + \frac{\sin 3\omega t}{3} + \frac{\sin 5\omega t}{5} + \cdots\right] \qquad [182]$$

which checks Eq. 172 for $x = \omega t$. The amplitude and phase functions are

$$|\alpha_n| = \frac{2}{\pi n} \qquad \text{(for } n \text{ odd)} \qquad [183]$$

which may be written

$$|\alpha(n\omega)| = \frac{2\omega}{\pi} \cdot \frac{1}{n\omega} \qquad \text{(for } n \text{ odd)} \qquad [184]$$

and

$$\phi_n = -\frac{\pi}{2} \qquad [185]$$

Since the phase function is a constant, there is no need to plot it. The plot of the amplitude function 183 or 184 is facilitated through plotting first a continuous dotted curve for this function, assuming $n\omega$ to be a continuous variable, and then erecting a set of ordinates from the $n\omega$-axis to this dotted curve at the points corresponding to odd integer values of n, as is illustrated in Fig. 11. The dotted curve in this example is simply a rectangular hyperbola. After this dotted curve is drawn, any number of lines are readily inserted. A plot of this sort gives one a good idea of the rate of convergence of the series, since the ordinates are proportional to the values of the various harmonic amplitudes.

A second interesting example is given by the function shown in Fig. 12. This consists of a periodic succession of identical rectangular pulses of

unit amplitude and duration δ, with the time origin chosen at the center of one of them. Application of the formula 176 in this case yields

$$\alpha_n = \frac{\omega}{2\pi}\int_{-\delta/2}^{\delta/2} e^{-jn\omega t}\,dt = \left.\frac{e^{-jn\omega t}}{-2\pi nj}\right]_{-\delta/2}^{\delta/2} = \frac{\sin n\omega\dfrac{\delta}{2}}{n\pi} \qquad [186]$$

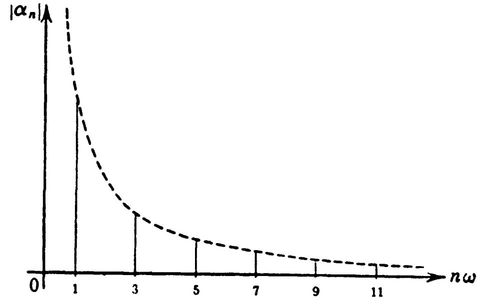


FIG. 11. The amplitude function in the Fourier analysis of a rectangular wave.

which is preferably written

$$\alpha_n = \frac{\delta\omega}{2\pi}\cdot\frac{\sin n\omega\dfrac{\delta}{2}}{n\omega\dfrac{\delta}{2}} = \frac{\delta}{\tau}\cdot\frac{\sin n\omega\dfrac{\delta}{2}}{n\omega\dfrac{\delta}{2}} \qquad [187]$$

Since the expression is real, the phase function in this example is zero, as is to be expected from the fact that the function $f(t)$ is even.

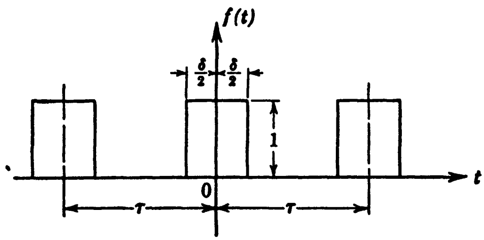


FIG. 12. A periodic succession of rectangular pulses.

The amplitude function has the form $(\sin x)/x$ which is equal to unity at $x = 0$ and zero at integer multiples of π. The general appearance of this function is that of a harmonic oscillation of decreasing amplitude,

although it is, of course, not simple harmonic. At the point $x = \pi/2$ the value is $2/\pi$; at $x = 3\pi/2$ it is $(1/3)(2/\pi)$; at $x = 5\pi/2$ it is $(1/5)(2/\pi)$, and so forth, the maxima and minima lying approximately at the points $x = 3\pi/2, 5\pi/2, \cdots$. With these items in mind a dotted curve showing the function 187 versus $n\omega$ is readily plotted. It remains to draw in the ordinates corresponding to integer values of n. Before this can be done one must choose a particular value for the ratio of the duration δ of the impulse to the fundamental period τ. For the present example, this ratio is chosen as

$$\frac{\delta}{\tau} = \frac{1}{5} \tag{188}$$

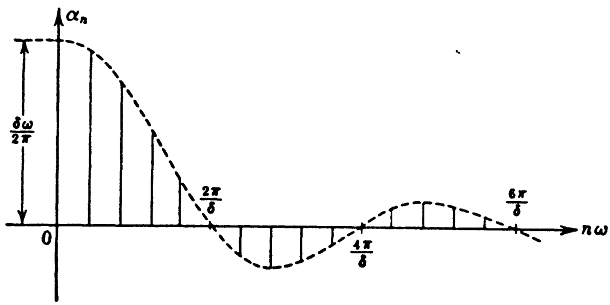


FIG. 13. The amplitude function in the Fourier analysis of the succession of rectangular pulses of Fig. 12.

and since $\tau = 2\pi/\omega$, this choice corresponds to a fundamental angular frequency of

$$\omega = \frac{1}{5}\cdot\frac{2\pi}{\delta} \tag{189}$$

As shown in Fig. 13, the point $n\omega = 2\pi/\delta$ marks the first zero of the dotted curve, and since the fundamental frequency corresponds to $n = 1$, this point is located one-fifth the distance from the origin to the point $n\omega = 2\pi/\delta$. The various ordinates for other integer values of n are then readily drawn. For the particular choice indicated in Eq. 188, the amplitudes of the fifth, tenth, fifteenth, $\cdots$ harmonics are zero because these points coincide with the zeros of the dotted curve.

It is to be observed that the form of the dotted curve is the same regardless of the ratio δ/τ, and that only the spacing of the lines in the spectrum is dependent upon this value. Incidentally, the amplitude of the constant component, corresponding to $n = 0$, is seen to be the largest. All amplitudes are proportional to the duration δ, so that as the ratio

δ/τ is decreased, τ remaining constant, the dotted curve remains intact except for a change in the scale factor for its ordinates.

10. Power products and effective values

In the discussion of problems in electric circuit theory, determining the average value of the product of two periodic functions is at times necessary. Since this product is again a periodic function with no larger fundamental period, it is sufficient to restrict the evaluation to one period. If the two functions are denoted by $f_1(t)$ and $f_2(t)$, the problem calls for the evaluation of the integral

$$J = \frac{\omega}{2\pi}\int_0^{2\pi/\omega} f_1(t)f_2(t)\,dt \tag{190}$$

This evaluation is particularly simple if the complex form is used for the Fourier series representation of the functions $f_1(t)$ and $f_2(t)$. These functions may be expressed as

$$f_1(t) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jn\omega t} \tag{191}$$

and

$$f_2(t) = \sum_{m=-\infty}^{\infty} \beta_m e^{jm\omega t} \tag{192}$$

The product is given by the double sum

$$f_1(t)\cdot f_2(t) = \sum_{n,\,m=-\infty}^{\infty} \alpha_n\beta_m e^{j(n+m)\omega t} \tag{193}$$

Substituting this expression into the integral 190, and assuming that the series 191 and 192 are uniformly convergent so that the integration may be carried out term by term, one has

$$J = \frac{\omega}{2\pi}\sum_{n,\,m=-\infty}^{\infty} \alpha_n\beta_m \int_0^{2\pi/\omega} e^{j(n+m)\omega t}\,dt \tag{194}$$

The integral involved here is a very simple one. Its evaluation reads

$$\int_0^{2\pi/\omega} e^{j(n+m)\omega t}\,dt = \frac{e^{j(n+m)2\pi} - 1}{j(n+m)\omega} = \begin{cases} \dfrac{2\pi}{\omega} & \text{for } m = -n \\ 0 & \text{for } m \neq -n \end{cases} \tag{195}$$

Consequently all the terms in the double sum 194 vanish except those for which $m = -n$. The function J, however, still requires an infinite sum for its representation, that is,

$$J = \sum_{n=-\infty}^{\infty} \alpha_n\beta_{-n} \tag{196}$$

which is the desired result. It may be written in the alternative form

$$J = \alpha_0\beta_0 + \sum_{n=1}^{\infty} (\alpha_n\beta_{-n} + \alpha_{-n}\beta_n) \qquad [197]$$

or, if one writes,

$$\alpha_n = |\alpha_n| \, e^{j\phi_n} \qquad \beta_n = |\beta_n| \, e^{j\psi_n} \qquad [198]$$

the result 197 is found to be equivalent to

$$J = \alpha_0\beta_0 + 2 \sum_{n=1}^{\infty} |\alpha_n\beta_n| \cos (\psi_n - \phi_n) \qquad [199]$$

Because the integral 195 is zero for $m \neq -n$, the interesting result follows that the average value 190, given by 199, depends only upon the products of the amplitudes of harmonic components of *like* order. In other words, none of the cross-product terms resulting from the product of the two series for $f_1(t)$ and $f_2(t)$ contributes to the average value of this product. In a physical problem in which $f_1(t)$ represents a source voltage and $f_2(t)$ the resulting source current, J represents the average power delivered by the source. The integral 190 is, therefore, also referred to as the *power product* of two periodic functions. The coefficient $\cos (\psi_n - \phi_n)$, which occurs in Eq. 199, is in electric circuit problems called the *power factor* corresponding to the harmonic of nth order.

A very closely related problem is the evaluation of the integral

$$J' = \frac{\omega}{2\pi}\int_0^{2\pi/\omega} [f(t)]^2 \, dt \qquad [200]$$

which is referred to as the *mean square* value of the single periodic function $f(t)$. This is, however, merely the integral 190 for $f_1(t) = f_2(t) = f(t)$, and hence the desired result is expressed by Eq. 196 or 199 for $\alpha_n = \beta_n$. Hence, if $f(t)$ is assumed to be given by Eq. 191,

$$J' = \alpha_0^2 + 2 \sum_{n=1}^{\infty} |\alpha_n|^2 \qquad [201]$$

The square root of this value is called the root-mean-square value or *effective value* of the periodic function $f(t)$. Recalling, according to Eqs. 94 and 135, that the harmonic amplitudes c_n in Eq. 95 are given by twice the magnitudes of the complex coefficients α_n, one sees that the effective value of the periodic function $f(t)$ is given by the expression

$$[f(t)]_{\text{effective}} = \sqrt{\frac{c_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty} c_n^2} \qquad [202]$$

which contains the particular result that the effective or root-mean-square value of a sinusoidal function equals its amplitude divided by $\sqrt{2}$.

11. Summation formulas

It is useful to note that the Fourier series for various particular functions may be used to obtain in closed form the values for numerous special forms of infinite series. For example, from the function $h(x)$ of Fig. 8 and its series 141, or from the function $f(x)$ of Fig. 9 and its series 172, one obtains by setting $x = \pi/2$ the result that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + - \cdots \qquad [203]$$

Similarly, since the series 174 for the triangular function of Fig. 10 is still uniformly convergent for $x = 0$, one has

$$\frac{\pi^2}{8} = 1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots \qquad [204]$$

The evaluation of many other infinite series may be obtained in this way.

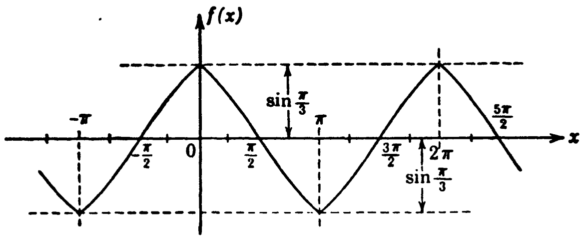


Fig. 14. A plot of the function of Eq. 205, the Fourier series of which is used in several summation formulas.

It is of greater practical value to recognize that similar means may be employed for the determination of more general summation formulas. For example, one may consider the function $f(x)$ defined by

$$\begin{aligned} f(x) &= \sin \mu\left(\frac{\pi}{2} + x\right) \quad (\text{for } -\pi < x < 0) \\ f(x) &= \sin \mu\left(\frac{\pi}{2} - x\right) \quad (\text{for } 0 < x < \pi) \end{aligned} \qquad [205]$$

in which μ may have any value. This function is shown plotted in Fig. 14 for the value $\mu = 2/3$. The function is even and satisfies the condition 98. Hence its Fourier series is represented by a cosine series with odd harmonics, and the coefficients are given by the formula 105. Substituting from Eq. 205, one has

$$a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx = \frac{2}{\pi}\int_0^{\pi} \sin \mu\left(\frac{\pi}{2} - x\right)\cos nx\,dx \qquad [206]$$

The integration yields

$$a_n = \left. \frac{2n \sin nx \sin \mu\left(\frac{\pi}{2} - x\right) - 2\mu \cos nx \cos \mu\left(\frac{\pi}{2} - x\right)}{\pi(n^2 - \mu^2)} \right]_0^\pi$$

$$= \frac{2\mu(1 - \cos n\pi) \cos \mu \frac{\pi}{2}}{\pi(n^2 - \mu^2)} \qquad [207]$$

or

$$a_n = \begin{cases} \dfrac{4\mu \cos \mu \frac{\pi}{2}}{\pi(n^2 - \mu^2)} & \text{(for } n \text{ odd)} \\ 0 & \text{(for } n \text{ even)} \end{cases} \qquad [208]$$

The Fourier series for this function is, therefore, given by

$$f(x) = \frac{4\mu \cos \mu \frac{\pi}{2}}{\pi} \sum_{n=1,3,5,\cdots}^{\infty} \frac{\cos nx}{n^2 - \mu^2} \qquad [209]$$

which converges uniformly for all values of x since the function $f(x)$ is continuous. Substituting from the relations 205, one may write for the interval $0 < x < \pi$

$$\frac{\pi \sin \mu\left(\frac{\pi}{2} - x\right)}{4\mu \cos \mu \frac{\pi}{2}} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\cos nx}{n^2 - \mu^2} \qquad [210]$$

Particularly, for $x = 0$, this result yields

$$\frac{\pi}{4\mu} \tan \mu \frac{\pi}{2} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{1}{(n^2 - \mu^2)} \qquad [211]$$

which is the partial fraction expansion of the tangent function, and is found to be a useful summation formula in connection with a number of problems in circuit theory.

Other formulas may be developed from the relations 210 and 211. For example, subtracting the former from the latter, and using the trigonometric identity

$$1 - \cos nx = 2 \sin^2 n \frac{x}{2} \qquad [212]$$

one finds

$$f(\mu,x) = \frac{\pi\left\{\sin\mu\frac{\pi}{2} - \sin\mu\left(\frac{\pi}{2} - x\right)\right\}}{8\mu\cos\mu\frac{\pi}{2}} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin^2 n\frac{x}{2}}{n(n^2-\mu^2)} \qquad [213]$$

which is a somewhat more general summation formula than Eq. 211.

Now writing the Eq. 213 first for $\mu = \mu_1$ and again for $\mu = \mu_2$, and subtracting the second from the first, give a more general formula which reads

$$\frac{f(\mu_1,x) - f(\mu_2,x)}{\mu_1^2 - \mu_2^2} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin^2 n\frac{x}{2}}{(n^2-\mu_1^2)(n^2-\mu_2^2)} \qquad [214]$$

The quantities μ, μ_1, and μ_2 in these formulas may have any real or complex values.

Formulas having the form of Eqs. 213 or 214 are useful in electric circuit theory because the quantities $(n^2 - \mu_k^2)$ have the form of the resonance factors of an impedance function. Since the latter is a rational function, that is, a quotient of polynomials, it is always factorable in terms of the roots of these polynomials. Hence the impedance is expressible as the quotient of products of factors of this type. The expression for the power absorbed by an electrical network which is excited by a periodic source containing only odd harmonics (this case is most common) assumes the form of the infinite sum in Eq. 214 when the network contains two degrees of freedom.

For $\mu_2 = 0$, a special form of Eq. 214 results, reading

$$\frac{8f(\mu,x) - \pi x}{8\mu^2} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin^2 n\frac{x}{2}}{n^2(n^2-\mu^2)} \qquad [215]$$

Other formulas may be obtained through differentiation of the preceding ones with respect to x. As long as the coefficients in the resulting series decrease for large values of n as rapidly as, or more rapidly than, the ratio $1/n^2$, then, according to Art. 8, the periodic function represented by the series is continuous, and the series still converges uniformly for all values of x inclusive of the boundary values in the relations 205. However, if the coefficients in the resultant series decrease for large values of n only as fast as the ratio $1/n$, the series cannot be used for the boundary values of x, although it still converges uniformly for all other

x-values. For example, differentiating the formula 213 with respect to x gives

$$\frac{\pi \cos \mu\left(\frac{\pi}{2} - x\right)}{4 \cos \mu \frac{\pi}{2}} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{n \sin nx}{(n^2 - \mu^2)} \qquad [216]$$

which is evidently no longer valid for $x = 0$ or $x = \pi$. However, differentiating the formula 215 with respect to x yields

$$\frac{\pi \cos \mu\left(\frac{\pi}{2} - x\right) - \pi \cos \mu \frac{\pi}{2}}{4\mu^2 \cos \mu \frac{\pi}{2}} = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin nx}{n(n^2 - \mu^2)} \qquad [217]$$

which converges uniformly for all x-values.

The formula 214 may be generalized so as to contain any number of factors $(n^2 - \mu_k^2)$ in the sum. Replacing μ_1 and μ_2 respectively by μ_2 and μ_3 and subtracting the resulting formula from 214 yield

$$\frac{f(\mu_1,x)}{(\mu_1^2-\mu_2^2)(\mu_1^2-\mu_3^2)} + \frac{f(\mu_2,x)}{(\mu_2^2-\mu_1^2)(\mu_2^2-\mu_3^2)} + \frac{f(\mu_3,x)}{(\mu_3^2-\mu_1^2)(\mu_3^2-\mu_2^2)}$$

$$= \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin^2 n \frac{x}{2}}{(n^2-\mu_1^2)(n^2-\mu_2^2)(n^2-\mu_3^2)} \qquad [218]$$

Again, replacing μ_1, μ_2, μ_3 in this formula by μ_2, μ_3, μ_4 respectively and subtracting the resulting formula from 218, one finds

$$\sum_{k=1}^{4} \frac{f(\mu_k,x)}{(\mu_k^2 - \mu_{k+1}^2)(\mu_k^2 - \mu_{k+2}^2)(\mu_k^2 - \mu_{k+3}^2)}$$

$$= \sum_{n=1,3,5,\cdots}^{\infty} \frac{\sin^2 n \frac{x}{2}}{(n^2 - \mu_1^2)(n^2 - \mu_2^2)(n^2 - \mu_3^2)(n^2 - \mu_4^2)} \qquad [219]$$

in which the subscripts $k = 1, 2, 3, 4$ are assumed to form a cyclic group.

Any desired further generalization is evidently possible. For example, by using Eq. 213, and forming the function $af(\mu_1,x) + bf(\mu_2,x)$, one obtains a formula which differs from Eq. 214 in that the terms in the sum contain a factor of the form $(n^2 - \mu_\nu^2)$ in the numerator in addition to factors of this form in the denominator.

12. The "least squares" approximation property of the Fourier series

In this article the problem of finding a trigonometric series representation for a given function $f(x)$ is approached in a different manner. The trigonometric series which is to approximate the function $f(x)$ is assumed to be finite and is written in the complex form

$$s_n(x) = \sum_{k=-n}^{n} a_k e^{jkx} \tag{220}$$

The following question is now raised: What must be the values of the coefficients a_k in order that the mean square error between $f(x)$ and $s_n(x)$ may become a minimum? In analytic form the mean square error, expressed for the fundamental range, reads

$$E = \frac{1}{2\pi}\int_a^{a+2\pi} [f(x) - s_n(x)]^2\, dx \tag{221}$$

The problem, therefore, is to determine the coefficients in Eq. 220 in such a way that the expression 221 becomes a minimum.

Now

$$[f(x) - s_n(x)]^2 = [f(x)]^2 + [s_n(x)]^2 - 2[f(x)s_n(x)] \tag{222}$$

According to Eq. 220,

$$[s_n(x)]^2 = \sum_{i,k=-n}^{n} a_i a_k e^{j(i+k)x} \tag{223}$$

so that

$$\frac{1}{2\pi}\int_a^{a+2\pi} [s_n(x)]^2\, dx = \frac{1}{2\pi}\sum_{i,\,k=-n}^{n} a_i a_k \int_a^{a+2\pi} e^{j(i+k)x}\, dx \tag{224}$$

The integral appearing here has the same form as that in Eq. 195 of Art. 10. Hence

$$\frac{1}{2\pi}\int_a^{a+2\pi} [s_n(x)]^2\, dx = \sum_{k=-n}^{n} a_k a_{-k} \tag{225}$$

Next one needs to evaluate the integral

$$\frac{1}{2\pi}\int_a^{a+2\pi} f(x)s_n(x)\, dx = \sum_{k=-n}^{n} a_k \left\{\frac{1}{2\pi}\int_a^{a+2\pi} f(x)e^{jkx}\, dx\right\} \tag{226}$$

The quantity enclosed by the curved brackets is, according to Eq. 122 or Eq. 156, recognized to be the expression for the complex Fourier coefficient

α_k, so that

$$\frac{1}{2\pi}\int_a^{a+2\pi} f(x)s_n(x)\,dx = \sum_{k=-n}^{n} a_k\alpha_{-k} \tag{227}$$

Substituting Eq. 222 into Eq. 221, and taking note of the results stated in Eqs. 225 and 227, one obtains for the mean square error the expression

$$E = \frac{1}{2\pi}\int_a^{a+2\pi} [f(x)]^2\,dx + \sum_{k=-n}^{n} a_k a_{-k} - 2\sum_{k=-n}^{n} a_k\alpha_{-k} \tag{228}$$

Now it is observed that

$$2\sum_{k=-n}^{n} a_k\alpha_{-k} = \sum_{k=-n}^{n} (a_k\alpha_{-k} + a_{-k}\alpha_k) \tag{229}$$

The last two terms in Eq. 228 may, therefore, be written in the form

$$\sum_{k=-n}^{n} (a_k a_{-k} - a_k\alpha_{-k} - a_{-k}\alpha_k + \alpha_k\alpha_{-k}) - \sum_{k=-n}^{n} \alpha_k\alpha_{-k} \tag{230}$$

which is equivalent to

$$\sum_{k=-n}^{n} (a_k - \alpha_k)(a_{-k} - \alpha_{-k}) - \sum_{k=-n}^{n} \alpha_k\alpha_{-k} \tag{231}$$

or, since the coefficients with negative subscripts are the conjugates of those with the same positive subscripts, these terms are given by

$$\sum_{k=-n} |a_k - \alpha_k|^2 - \sum_{k=-n}^{n} |\alpha_k|^2 \tag{232}$$

Hence the mean square error, according to Eq. 228, is seen to be given by

$$E = \frac{1}{2\pi}\int_a^{a+2\pi} [f(x)]^2\,dx + \sum_{k=-n}^{n} |a_k - \alpha_k|^2 - \sum_{k=-n}^{n} |\alpha_k|^2 \tag{233}$$

Inasmuch as the function $f(x)$ and hence also the coefficients α_k are fixed, the expression 233 evidently becomes a minimum for

$$a_k = \alpha_k \tag{234}$$

whence one may conclude that the sum $s_n(x)$ as given by Eq. 220 approximates the stated function $f(x)$ over its fundamental range so as to make the mean square error a minimum *if the coefficients in the finite sum* $s_n(x)$ *are the Fourier coefficients for the function* $f(x)$.

When the coefficients a_k are so determined, the mean square error 233 becomes

$$E = \frac{1}{2\pi}\int_a^{a+2\pi} [f(x)]^2\,dx - \sum_{k=-n}^{n} |\alpha_k|^2 \tag{235}$$

which, according to Art. 10, is

$$E = \sum_{k=-\infty}^{\infty} |\alpha_k|^2 - \sum_{k=-n}^{n} |\alpha_k|^2 \tag{236}$$

Thus it is seen that the mean square error tends toward zero as n becomes larger and larger. In other words, the infinite Fourier series approximates the given function $f(x)$ in such a way as to make the mean square error over the fundamental range vanishingly small. The Fourier series, or any of its partial sums, is, therefore, said to approximate a given function in the " least squares " sense.

In view of this result it is easy to show why, at a point of discontinuity of the function $f(x)$, the Fourier series always yields the arithmetic mean between the two values of $f(x)$ immediately adjacent to the point of discontinuity. Thus, if these two values are denoted by a and b, and the value of the series at the point of discontinuity is denoted by s, the mean square error at this point is expressed by

$$\frac{(s - a)^2 + (s - b)^2}{2} \tag{237}$$

If this is to be a minimum, its derivative with respect to s must be zero, that is,

$$s - a + s - b = 0 \tag{238}$$

which yields

$$s = \frac{a + b}{2} \tag{239}$$

13. The approximation property of the partial sums; the Gibbs phenomenon

When, in a practical problem, a trigonometric series is used to approximate a given function, this series must for obvious reasons be a finite one. Consequently, it is of considerable interest to know more about the detailed manner in which the partial sums of a Fourier series approximate a prescribed function.

A partial sum reads

$$s_n(x) = \sum_{k=-n}^{n} \alpha_k e^{jkx} \tag{240}$$

in which

$$\alpha_k = \frac{1}{2\pi} \int_0^{2\pi} f(\xi) e^{-jk\xi}\, d\xi \tag{241}$$

Substituting this expression for α_k into Eq. 240 and reversing the order of summation and integration yield

$$s_n(x) = \frac{1}{2\pi}\int_0^{2\pi} f(\xi) \sum_{k=-n}^{n} e^{jk(x-\xi)}\, d\xi \qquad [242]$$

Now

$$\tfrac{1}{2}\sum_{k=-n}^{n} e^{jk(x-\xi)} = \tfrac{1}{2} + \cos(x-\xi) + \cos 2(x-\xi) + \cdots + \cos n(x-\xi) \qquad [243]$$

and, according to Eq. 11, this gives

$$\sum_{k=-n}^{n} e^{jk(x-\xi)} = \frac{\sin\left\{(2n+1)\dfrac{x-\xi}{2}\right\}}{\sin\left(\dfrac{x-\xi}{2}\right)} \qquad [244]$$

Substituting this result into Eq. 242, one finds that a partial sum may be expressed as

$$s_n(x) = \frac{1}{2\pi}\int_0^{2\pi} f(\xi)u(\xi - x)\, d\xi \qquad [245]$$

in which

$$u(\xi - x) = \frac{\sin\left\{(2n+1)\dfrac{\xi - x}{2}\right\}}{\sin\left(\dfrac{\xi - x}{2}\right)} \qquad [246]$$

This function is shown plotted versus ξ in Fig. 15 for the particular choice $n = 5$. It consists of a periodic succession of identical large humps with the amplitude $(2n + 1)$ and the fundamental period 2π, separated by a number of smaller oscillations which decrease in amplitude with increasing distance from the large humps, the smallest of these having unit amplitude and occurring midway between any two of the large humps. It is observed that only one large hump is encountered throughout the fundamental range $0 \leqq x \leqq 2\pi$, and this occurs at the point $\xi = x$.

According to the relation 245, the partial sum $s_n(x)$ is given by the product of this peculiar function $u(\xi - x)$ with the given function $f(\xi)$ integrated over the fundamental range. The integration must be evaluated separately for each value of x for which the corresponding value of $s_n(x)$ is desired. During each integration, x is treated as a constant. The function $u(\xi - x)$, which contains x as a parameter, remains the same for various x-values except for a translation of this function as a whole.

The value of x merely determines the location of the large hump in the fundamental range.

Figure 16 illustrates the function $u(\xi - x)$ for a fairly large value of n, and shows its position in the fundamental range relative to the given function $f(\xi)$ for some arbitrary x-value.* For the interpretation of the integration in Eq. 245 one may visualize the plane of Fig. 16 as a screen on which the function $u(\xi - x)$ is plotted and of which the portion corresponding to the shaded area under this curve is cut out. The function $f(\xi)$, which is plotted on a second screen located behind the first, is

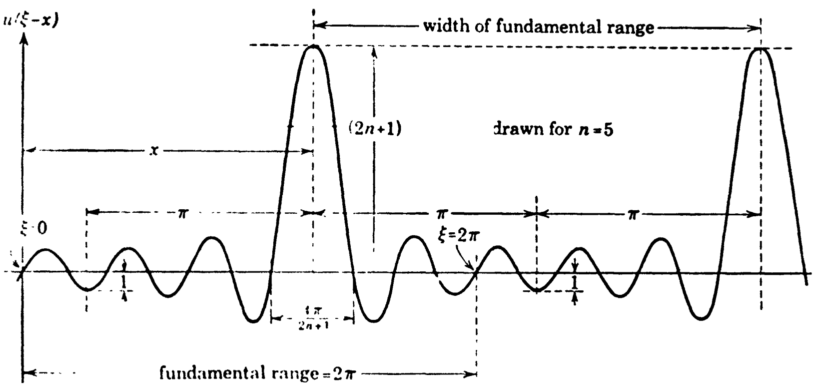


FIG. 15. The scanning function used in a study of the Gibbs phenomenon.

assumed to be visible only through the portion of the first screen which is cut away. For a large value of n, this cutaway portion is essentially in the form of a long narrow slit corresponding to the area under the large hump in the function $u(\xi - x)$. If one neglects for the moment the small areas under the minor oscillations or ripples adjacent to this slit, and assumes further that this slit is sufficiently narrow so that the small portion of the function $f(\xi)$ visible through it may to a first approximation be regarded as characterizing the single ordinate of this function for the point $\xi = x$, it becomes clear that the value of the integral 245 is given simply by the product of $f(x)$ and the area of the slit multiplied by $1\ 2\pi$. If $1/2\pi$ times the area of the slit equals unity, the value of the integral, to a good approximation, equals $f(x)$. In other words, one arrives at the reasonable conclusion that $s_n(x) \cong f(x)$.

*Since the ratio of the height of the large hump in the function $u(\xi - x)$ to the width of the fundamental range is $(2n + 1)/2\pi$, it is not practically feasible to draw this figure, or Fig. 20, in correct proportion with regard to this ratio. These figures are, however, in correct proportion with regard to the ratios of the amplitude of the large hump to those of the adjacent oscillations, as well as with regard to the ratio of the width of the fundamental range to that of the large hump at its base. This ratio is equal to $n + \frac{1}{2}$.

The values of $s_n(x)$ for various values of x are obtained through displacing the first screen in the horizontal direction so that the slit uncovers different portions of the curve $f(\xi)$ corresponding to the point $\xi = x$. This process of moving the screen with the slit corresponding to $u(\xi - x)$ across the fundamental range, and viewing the function $f(\xi)$ through it, is referred to as *scanning* $f(\xi)$ with the function $u(\xi - x)$. The latter is called the *scanning function*. If this scanning function consists of a single rectangular hump of extremely small width but of such height that the area of the slit nevertheless is finite and equal to 2π, the integral 245 yields almost exactly the function $f(x)$, the discrepancy becoming zero

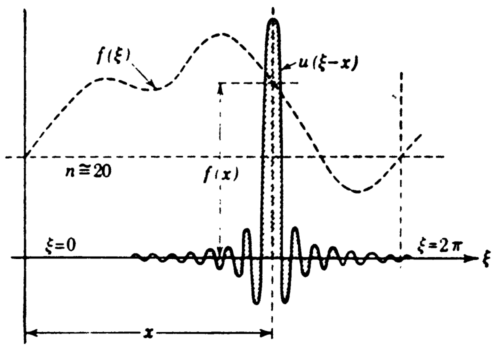


FIG. 16. Pertinent to the evaluation of the partial sum of a Fourier series.

as the slit width is allowed to approach zero. Actually, because of the deviation of the scanning function $u(\xi - x)$ from this ideal limiting form, the integral yields a function $s_n(x)$ which is only approximately equal to $f(x)$.

For a relatively small value of n (few terms of the Fourier series) the scanning function $u(\xi - x)$ deviates quite considerably from its ideal form, and the partial sum $s_n(x)$, as one should expect, only crudely approximates the given function $f(x)$. As more terms are added to the partial sum, and n becomes relatively large, the form of the scanning function improves, and so does the degree of approximation between $s_n(x)$ and $f(x)$. The improvement in the scanning function is due to the large hump's becoming taller and narrower, as is clear from Fig. 15, in which it is indicated that the height of the large hump equals $(2n + 1)$ and its width at the base equals $4\pi/(2n + 1)$. At the same time, the adjacent ripples become larger in number and hence also become narrower, in the same proportion, in fact, as the large hump becomes narrower.

One disturbing feature about the behavior of this scanning function with increasing n, however, is that the ripples immediately adjacent to

the large hump do not ultimately become smaller and smaller in amplitude relative to the height of the large hump but maintain amplitudes which asymptotically approach a finite ratio relative to the height of the large hump. The ultimate effect of this disturbing circumstance requires further careful investigation which is begun by study of the properties of the scanning function in greater detail.

First it is significant to observe that the net area under the curve for $u(\xi - x)$ over the fundamental range is indeed equal to 2π, and that this is true for any n, as may readily be seen in several ways. For example, according to Eqs. 243, 244, and 246,

$$u(\xi-x)=1+2\cos(\xi-x)+2\cos 2(\xi-x)+\cdots+2\cos n(\xi-x) \quad [247]$$

In evaluating the integral of this expression over the range $0 < \xi < 2\pi$ it is observed that all the integrals for the cosine terms yield zero. Hence

$$\int_0^{2\pi} u(\xi - x)\,d\xi = \int_0^{2\pi} d\xi = 2\pi \quad [248]$$

This result may also be seen from the integral 245 through assuming as a special example that $f(\xi) = 1$. Then $s_n(x)$ must, of course, also equal unity; that is, the partial sum reduces to its constant term and this must have the value unity. The integral 245 then yields the result expressed by Eq. 248.

Thus it is established that $1/2\pi$ times the area under the scanning function is equal to unity, as the preceding discussion indicates that it should be. However, this area is not confined to a single slit but is partly contributed to by smaller increments due to the ripples, which are alternately numerically positive and negative. A glance at the scanning functions shown in Figs. 15 and 16 reveals, moreover, that the net area contributed by the ripples is numerically negative so that the area under the large hump must exceed the value 2π if the resultant area under the entire function is to have this value. The amount of the excess area under the large hump depends upon the value of n, but the important point is that this excess does not become zero as n is indefinitely increased.

In order to comprehend this fact, one must study the behavior of the function $u(\xi - x)$ for large values of n. To facilitate this study it is expedient to introduce in the form given by Eq. 246 the change of variable

$$\eta = (2n + 1)\frac{\xi - x}{2} \quad [249]$$

Then the scanning function assumes the form

$$u(\eta) = \frac{\sin \eta}{\sin\left(\dfrac{\eta}{2n + 1}\right)} \quad [250]$$

The region of the main hump corresponds to $-\pi < \eta < \pi$, and each adjacent smaller hump corresponds to an increment of π in the variable η. As indicated in Fig. 16, the region over which the adjacent ripples have significant amplitudes is confined to the more immediate vicinity of the large hump. In other words, the entire region of interest for the function $u(\eta)$ is confined to a finite interval for the variable η which, for example, may be described as $-10\pi < \eta < 10\pi$, or something of this order.

With regard to the expression 250, if n is sufficiently large so that approximately

$$\frac{10\pi}{2n+1} < \frac{\pi}{6} \tag{251}$$

little error is introduced in the determination of $u(\eta)$ throughout the region of interest if the sine function in the denominator is replaced by its argument. In the limit $n \to \infty$ any error introduced by this procedure

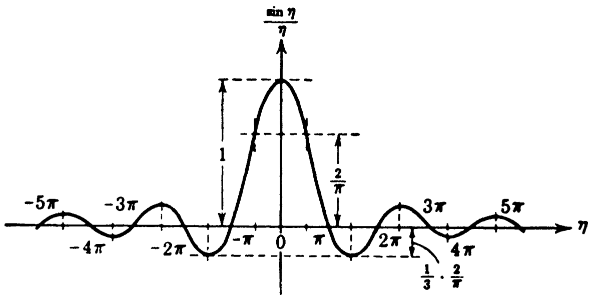


FIG. 17. A plot of the function $\dfrac{\sin\eta}{\eta}$ which occurs many times in Fourier analysis.

becomes vanishingly small. Hence, for large values of n, one is permitted to write for the scanning function

$$u(\eta) = (2n+1)\left(\frac{\sin\eta}{\eta}\right) \tag{252}$$

The function $(\sin\eta)/\eta$ has appeared before in these discussions (see Fig. 13 of Art. 9) and will appear again in a variety of problems. The reader may, therefore, well spend some time becoming thoroughly familiar with its characteristics. It is plotted again in Fig. 17, which illustrates some of its more immediately apparent properties. At the origin, $\eta = 0$, the function has the value unity, and a zero derivative,

as is apparent if one writes the Maclaurin series for $(\sin \eta)/\eta$, thus

$$\frac{\sin \eta}{\eta} = \frac{\eta - \dfrac{\eta^3}{3!} + \dfrac{\eta^5}{5!} - \cdots}{\eta} = 1 - \frac{\eta^2}{6} + \frac{\eta^4}{120} - + \cdots \qquad [253]$$

from which it is also clear that the function is even. At the points $\eta = \pm\pi, \pm 2\pi, \cdots$ the function passes through zero. Aside from the maximum at $\eta = 0$, the function has further maxima and minima at the points where

$$\frac{d}{d\eta}\left(\frac{\sin \eta}{\eta}\right) = 0 \qquad [254]$$

or for

$$\tan \eta = \eta \qquad [255]$$

The roots of this transcendental equation are very nearly, and for sufficiently large η almost exactly, equal to

$$\eta = \frac{3\pi}{2}, \quad \frac{5\pi}{2}, \quad \frac{7\pi}{2}, \cdots \qquad [256]$$

Actually they are slightly smaller than these values, but as a sketch of the relation 255 readily reveals, the discrepancies even for the first few roots in the set 256 are hardly noticeable in a graphical representation. The values of these approximate maxima and minima are most conveniently expressed in terms of the value of the function for $\eta = \pm\pi/2$, which is $2/\pi$. The approximate minimum at $\eta = \pm 3\pi/2$, which is a negative value, is one third of $2/\pi$; the maximum at $\eta = \pm 5\pi/2$ is one fifth of $2/\pi$; and so forth. In other words, the maxima and minima in absolute value become smaller with increasing η like the ratio $1/\eta$. For example, the fourth extremum at $7\pi/2$ has the value $2/7\pi = 0.091$, or only about 9 per cent of the value of the function at $\eta = 0$.

With regard to the area under the curve of Fig. 17 it is significant that the function

$$Si\,(x) = \int_0^x \frac{\sin \eta}{\eta}\, d\eta \qquad [257]$$

which represents the area under the curve for $(\sin \eta)/\eta$ from the origin to some variable point $\eta = x$ is a familiar function in Fourier analysis and allied matters. A short discussion of some of its properties may well be given at this time.

The shaded area in Fig. 18 represents the value of the function $Si\,(x)$, called the "sine-integral of x," for a particular value of x. The essential

properties of this function are easily obtained from inspection of this figure.

First one may observe that for values of x near the origin, the area evidently increases very nearly, in fact for $x = 0$ exactly, as x. This circumstance may readily be seen by observing that the function $(\sin \eta)/\eta$ has a zero derivative (the curve is horizontal) at the origin. Hence

$$\left[\frac{d\, Si\,(x)}{dx}\right]_{x=0} = 1 \qquad [258]$$

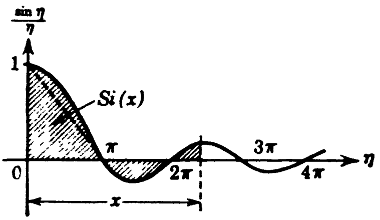


FIG. 18. The shaded area is the sine-integral of x.

As the curve for $(\sin \eta)/\eta$ begins to fall, however, the rate at which $Si\,(x)$ increases with x falls off, and it eventually becomes zero at $x = \pi$. Here the function $Si\,(x)$ reaches its first maximum. From $x = \pi$ to $x = 2\pi$ the Si-function decreases from its first maximum by the amount of the negative area enclosed by this portion of the $(\sin \eta)/\eta$ curve. By

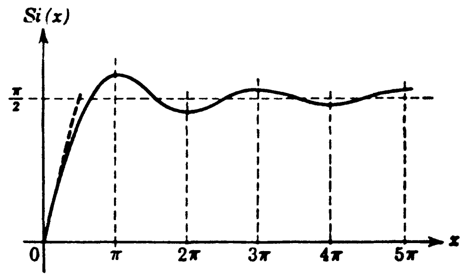


FIG. 19. A plot of the sine-integral of x.

the same sort of reasoning, it is readily seen that the function $Si\,(x)$ has the general character shown in Fig. 19. The alternately positive and negative areas contributed by the ripples in the function $(\sin \eta)/\eta$ become smaller and smaller with increasing x values, and the net area, therefore, approaches a finite value asymptotically and in an oscillatory manner.

The asymptotic value is readily obtained from the result stated by Eq. 248, namely, that the net area under the curve for the scanning function is equal to 2π for any n and any x. This equation may be written

$$\int_{-\pi}^{\pi} u(\xi - x)d(\xi - x) = 2\pi \qquad [259]$$

or, with the use of Eqs. 249 and 250, one has

$$\frac{2}{2n+1}\int_{-(2n+1)\pi/2}^{(2n+1)\pi/2} \frac{\sin \eta}{\sin\left(\dfrac{\eta}{2n+1}\right)}\, d\eta = 2\pi \tag{260}$$

In the limit $n \to \infty$ this result reads

$$2\int_{-\infty}^{\infty} \frac{\sin \eta}{\eta}\, d\eta = 2\pi \tag{261}$$

and since the function $(\sin \eta)/\eta$ is even, it follows that

$$\int_{0}^{\infty} \frac{\sin \eta}{\eta}\, d\eta = \frac{\pi}{2} \tag{262}$$

which is the asymptotic value of the function $Si\ (x)$, as indicated in Fig. 19.

With regard to the evaluation of the integral 245 for large values of n, according to the scanning process illustrated in Fig. 16, several points of significance may now be brought to the reader's attention. First it is to be observed from Fig. 17, which according to Eq. 252 represents the appearance of the scanning function for large values of n, that the smaller negative humps immediately adjacent to the large hump have amplitudes which remain slightly greater than one-fifth the amplitude of the large hump no matter how large n becomes. As already pointed out, the scanning function, therefore, does not approach its ideal form as n increases without limit, since the net area under the scanning function does not become equal to the area under the large hump. The latter area is, for large values of n, equal to twice the area under the curve of Fig. 17 between $\eta = -\pi$ and $\eta = \pi$. (The factor of *two* enters here because of the change of variable given by Eq. 249, as is also seen in the steps leading to Eq. 261.) This area is equal to $4\,Si\,(\pi)$, which is found (from tables of Si-functions) to be about 18 per cent larger than 2π.

In the limit $n \to \infty$, the scanning function may be visualized as resulting when all the side oscillations in Fig. 16, together with the large hump, are compressed so as to form a single ordinate of infinite height. The excess area under the large hump then virtually becomes coincident with the residual areas of the ripples, and since the net area always equals 2π, the ultimate form of the scanning function may in a sense be said to have met the ideal requirements. Yet for any finite n, however large, there exists a departure from the ideal which has a definite effect upon the approximation property of the partial sum $s_n(x)$. This departure becomes particularly marked in examples in which the given function $f(x)$ possesses discontinuities.

In order to illustrate this fact a function $f(x)$ is assumed to be zero from $x = 0$ to $x = x_1$, and equal to unity over the remainder of the fundamental range from $x = x_1$ to $x = 2\pi$. The integral 245 then becomes

$$s_n(x) = \frac{1}{2\pi}\int_{x_1}^{2\pi} u(\xi - x)\,d\xi \qquad [263]$$

This expression may be replaced by the two integrals

$$s_n(x) = \frac{1}{2\pi}\int_{x}^{2\pi} u(\xi - x)\,d\xi - \frac{1}{2\pi}\int_{x}^{x_1} u(\xi - x)\,d\xi \qquad [264]$$

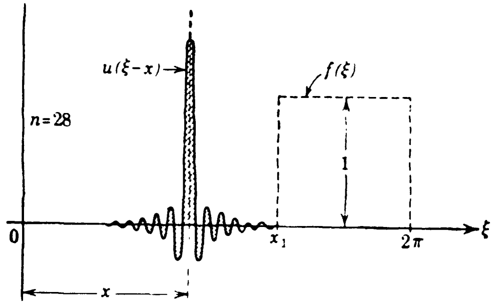


FIG. 20. The scanning function for $n = 28$ applied to a rectangular pulse.

These manipulations are clarified by reference to Fig. 20, in which the scanning function for $n = 28$ as well as the particular function $f(\xi)$ considered in this example is plotted. Thus it becomes clear that the first integral represents $1/2\pi$ times the area under half the scanning function and, therefore, has the value $\frac{1}{2}$. By use of Eqs. 249 and 252 in the second integral, the relation 264 becomes

$$s_n(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^{(2n+1)(x_1-x)/2} \frac{\sin\eta}{\eta}\,d\eta \qquad [265]$$

which, according to Eq. 257 is equivalent to

$$s_n(x) = \frac{1}{2} + \frac{1}{\pi}\,Si\,[(n + \tfrac{1}{2})(x - x_1)] \qquad [266]$$

This expression shows how the discontinuity of the function $f(x)$ at $x = x_1$ is represented by the partial sum $s_n(x)$. The second discontinuity at $x = 2\pi$ is represented in an exactly similar manner, with the result that the partial sum for this periodic function $f(x)$ and the choice $n = 28$ assumes the form shown in Fig. 21.

It must be admitted, of course, that a discontinuity in the function $f(x)$ offers an extreme test of the ability of the partial sum $s_n(x)$ to approxi-

mate this function. In other words, for a continuous function $f(x)$, the partial sum $s_n(x)$ with the same number of terms yields a vastly better degree of approximation. Nevertheless, it is practically significant to investigate the approximation property of the partial sum in the most adverse circumstances. In this respect one observes from the appearance of Fig. 21 that residual discrepancies remain even for very large values of n. As the latter is further increased, this figure is changed only in that the ripples in the vicinity of the discontinuities of $f(x)$ show a proportionately increased rate of oscillation versus the variable x, *whereas their amplitudes relative to the magnitude of the discontinuity remain the same.*

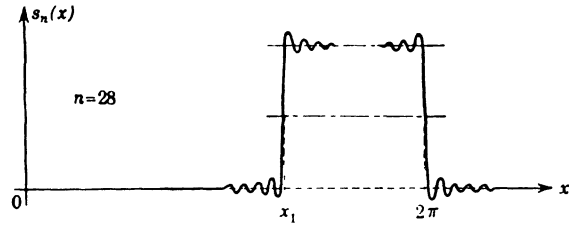


FIG. 21. The result of the operation illustrated in Fig. 20.

In the limit $n \to \infty$, these ripples are compressed into a single vertical line at the point of discontinuity, but even in this limit the Fourier series is still observed to yield the overswing of 18 per cent* which is characteristic of the function $Si(x)$. It is true, of course, that in the limit $n \to \infty$ this overswing together with the adjacent ripples occupies zero space in the fundamental range, so that practically speaking one may say that no residual discrepancy between the series and the function remains. Nevertheless, the phenomenon is noteworthy from a mathematical standpoint inasmuch as it illustrates the ultimate effect of the failure of the scanning function properly to approach its ideal form in the limit $n \to \infty$. It is, moreover, of practical concern also since it reveals the disheartening fact that, by means of a finite portion of a Fourier series, a given function can never be approximated in the vicinity of a discontinuity with a tolerance less than the characteristic 18 per cent overswing, no matter how many terms one may be willing to use.

This peculiarity of the Fourier series, referred to in mathematical terms as the Gibbs phenomenon, is a very real disadvantage with regard to certain practical problems. Because of this phenomenon, it is necessary in connection with certain approximation problems to use other types of trigonometric series which are appropriate modifications of the Fourier

*This percentage is based upon half the value of the discontinuity. In terms of the whole discontinuity the overswing is 9 per cent.

series. An important modification of this sort is discussed in the next article.

14. Approximations by means of Fejér polynomials

The appearance of the Gibbs phenomenon in the Fourier series is evidence of the failure of the uniform convergence of that series in the vicinity of a discontinuity of the given function. In other words, the partial sums $s_n(x)$ no longer converge toward a definite limit with increasing n for the immediate vicinity of a point of discontinuity of $f(x)$. It is the object of the present article to show that this failure in the uniform convergence of the series, and the associated Gibbs phenomenon, may be removed if the sequence of the partial sums $s_n(x)$ is replaced by the arithmetic mean sequence defined by Eq. 146 of Art. 9, Ch. VI.

This sequence reads

$$\begin{aligned} S_0(x) &= s_0(x) \\ S_1(x) &= \frac{s_0(x) + s_1(x)}{2} \\ S_2(x) &= \frac{s_0(x) + s_1(x) + s_2(x)}{3} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ S_\nu(x) &= \frac{1}{\nu + 1} \sum_{n=0}^{\nu} s_n(x) \end{aligned} \tag{267}$$

Substituting from Eq. 240, and utilizing the notation

$$\alpha_{\pm k} = \frac{c_k}{2} e^{\pm j\phi_k} \tag{268}$$

one recognizes that this sequence may alternatively be expressed as

$$\begin{aligned} S_0(x) &= \alpha_0 = \frac{c_0}{2} \\ S_1(x) &= \alpha_0 + \tfrac{1}{2}c_1 \cos(x + \phi_1) \\ S_2(x) &= \alpha_0 + \tfrac{2}{3}c_1 \cos(x + \phi_1) + \tfrac{1}{3}c_2 \cos(2x + \phi_2) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ S_\nu(x) &= \alpha_0 + \frac{\nu - 1}{\nu} c_1 \cos(x + \phi_1) + \frac{\nu - 2}{\nu} c_2 \cos(2x + \phi_2) \\ &\quad + \cdots + \frac{1}{\nu} c_\nu \cos(\nu x + \phi_\nu) \end{aligned} \tag{269}$$

The coefficients α_k are the Fourier coefficients defined by Eq. 241, and hence the partial sums appearing in Eq. 267 are, according to Eqs. 245 and 246, given by

$$s_n(x) = \frac{1}{2\pi}\int_0^{2\pi} f(\xi)\cdot\frac{\sin\left\{(2n+1)\frac{\xi - x}{2}\right\}}{\sin\left\{\frac{\xi - x}{2}\right\}}\cdot d\xi \qquad [270]$$

For the following manipulations it is now useful to observe that

$$\frac{\sin\left\{(2n+1)\frac{\xi - x}{2}\right\}}{\sin\left(\frac{\xi - x}{2}\right)} = \frac{\sin\left\{(2n+1)\frac{\xi - x}{2}\right\}\sin\left(\frac{\xi - x}{2}\right)}{\sin^2\left(\frac{\xi - x}{2}\right)}$$
$$= \frac{\cos n(\xi - x) - \cos[(n+1)(\xi - x)]}{2\sin^2\left(\frac{\xi - x}{2}\right)} \qquad [271]$$

By use of Eq. 270 in the last of the relations 267, it is found that

$$S_\nu(x) = \frac{1}{2\pi}\int_0^{2\pi} f(\xi)\sum_{n=0}^{\nu}\frac{\cos n(\xi - x) - \cos[(n+1)(\xi - x)]}{2(\nu+1)\sin^2\left(\frac{\xi - x}{2}\right)}\cdot d\xi \qquad [272]$$

But

$$\sum_{n=0}^{\nu}\{\cos n(\xi - x) - \cos[(n+1)(\xi - x)]\} = 1 - \cos(\xi - x) + \cos(\xi - x)$$
$$- \cos 2(\xi - x) + \cos 2(\xi - x) - \cdots - \cos(\nu+1)(\xi - x)$$
$$= 1 - \cos(\nu+1)(\xi - x) = 2\sin^2\left\{(\nu+1)\frac{\xi - x}{2}\right\} \qquad [273]$$

Hence Eq. 272 becomes

$$S_\nu(x) = \frac{1}{2\pi}\int_0^{2\pi} f(\xi)\cdot v(\xi - x)\cdot d\xi \qquad [274]$$

in which

$$v(\xi - x) = \frac{1}{\nu+1}\cdot\left(\frac{\sin\left\{(\nu+1)\frac{\xi - x}{2}\right\}}{\sin\left(\frac{\xi - x}{2}\right)}\right)^2 \qquad [275]$$

The expression 274 for the trigonometric polynomial $S_\nu(x)$ is identical

with the form 245 for the partial sum $s_n(x)$ except that the scanning function now is $v(\xi - x)$, as given by Eq. 275, instead of $u(\xi - x)$ as given by Eq. 246. Comparing the function $u(\xi - x)$ with $v(\xi - x)$ shows that the latter is essentially the square of the former. With $\nu = 2n$, this simple relationship between $u(\xi - x)$ and $v(\xi - x)$ is true except for the factor $1/(\nu + 1)$ in the expression for $v(\xi - x)$. But for this scale factor, the plot of Eq. 275 for $\nu = 10$ is simply given by the square of the function $u(\xi - x)$ plotted in Fig. 15 for $n = 5$. The result is illustrated in Fig. 22.

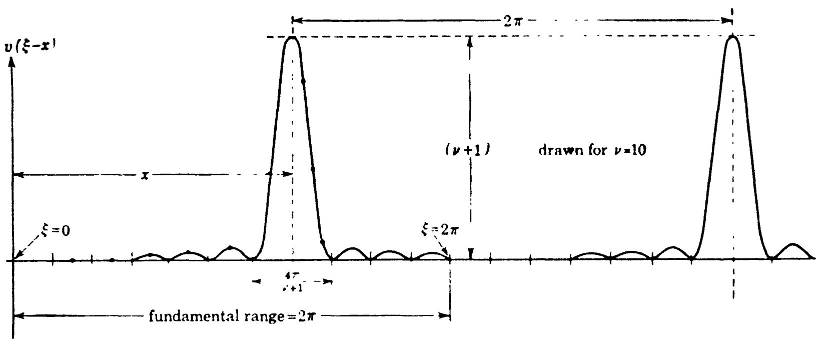


FIG. 22. A plot of the scanning function that results when Cesàro summation is applied to a Fourier series.

A comparison of Figs. 15 and 22 clearly reveals the superiority of the scanning function $v(\xi - x)$ over the function $u(\xi - x)$. Most significant in this respect is that the present function has no negative values, from which it follows that, at a point of discontinuity of the function $f(x)$, the partial sum $S_\nu(x)$ cannot overswing. Moreover, the ripples adjacent to the large hump in the function $v(\xi - x)$ have considerably smaller amplitudes than those for the function $u(\xi - x)$.

The net area under the scanning function $v(\xi - x)$ for the fundamental range is again equal to 2π, as may most easily be seen through considering the integral 274 for the special case $f(\xi) = 1$, whence $S_\nu(x)$ must also equal unity. Since the function $v(\xi - x)$ has the period 2π, it is thus established that

$$\frac{1}{2\pi(\nu + 1)} \int_{-\pi}^{\pi} \left(\frac{\sin\left\{(\nu + 1)\dfrac{\xi - x}{2}\right\}}{\sin\left(\dfrac{\xi - x}{2}\right)} \right)^2 d\xi = 1 \qquad [276]$$

for any x and any ν.

With the change of variable

$$\eta = (\nu + 1)\frac{\xi - x}{2} \tag{277}$$

the function 275 becomes

$$v(\eta) = \frac{1}{\nu + 1}\left(\frac{\sin \eta}{\sin \dfrac{\eta}{\nu + 1}}\right)^2 \tag{278}$$

For large values of ν the scanning function is, therefore, very nearly given by

$$v(\eta) = (\nu + 1)\cdot\left(\frac{\sin \eta}{\eta}\right)^2 \tag{279}$$

The integral 276 then reads

$$\frac{1}{\pi}\int_{-(\nu+1)\pi/2}^{(\nu+1)\pi/2}\left(\frac{\sin \eta}{\eta}\right)^2 d\eta = 1 \tag{280}$$

and for $\nu \longrightarrow \infty$ one has

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{\sin \eta}{\eta}\right)^2 d\eta = 1 \tag{281}$$

From Eq. 279 it is found for large ν-values that the area under the large hump in the scanning function $v(\xi - x)$ is equal to approximately 90 per cent of the total area of 2π, and that the combined area under the two small humps on either side is about 5 per cent of the total area. The amplitude of the small humps immediately adjacent to the large one is about 4.5 per cent of the large amplitude. These ratios hold for reasonably large finite values of ν (about 10 or more) and do not vary as ν is increased without limit.

With this scanning function, the function $f(\xi)$ of Fig. 20 yields a partial sum $S_\nu(x)$, according to Eq. 269, which has the form shown in Fig. 23. In order that this figure may correspond to the same width of the large hump and of the adjacent ripples relative to the fundamental range as for the scanning function of Fig. 20, it is drawn for $\nu = 56$, which is double the n-value chosen for Fig. 20. In other words, it is necessary to have twice as many terms in the partial sum 269 as in the partial sum for the Fourier series in order to obtain the same slope at the points of discontinuity of the function $f(x)$. However, the Gibbs phenomenon is now completely suppressed. Instead, the increment in $S_\nu(x)$ near the point x_1 equals only about 90 per cent of the value of the discontinuity, corresponding to a displacement of the scanning function $v(\xi - x)$ by an amount equal to the width at the base of its large hump. A succeeding

displacement equal to the width of the next adjacent small hump yields a further increase in $S_\nu(x)$ of about 5 per cent, and so on.

The essential difference between the behavior of the partial sum $S_\nu(x)$ and that of $s_n(x)$ in the vicinity of a discontinuity of the function

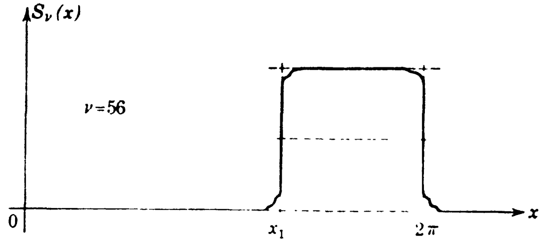


FIG. 23. The result of scanning a rectangular pulse with the scanning function of Fig. 22.

$f(x)$ may be seen from a comparison of the function $Si\ (x)$, defined by Eq. 257 and illustrated in Fig. 19, with the function

$$Q(x) = \int_0^x \left(\frac{\sin \eta}{\eta}\right)^2 d\eta \qquad [282]$$

The latter is shown plotted in Fig. 24, in which the dotted curve representing $Si\ (x)$ is also drawn in order to facilitate the comparison of these two functions. They both have the same asymptote, but $Si\ (x)$ converges toward this value in an oscillatory manner, whereas the function $Q(x)$ approaches it monotonically.

The conclusion is that the so-called *Fejér* sum $S_\nu(x)$, given by Eq. 269, converges toward a definite value as ν is indefinitely increased, even at points where the given function $f(x)$ is discontinuous. In other words, the Fejér series, which is given by $S_\nu(x)$ for $\nu \to \infty$, converges uniformly over the entire fundamental range even when the function $f(x)$ which it represents possesses discontinuities.

It is interesting to observe, according to the relations 269, that for large values of ν the coefficients in the sum $S_\nu(x)$ differ appreciably from those in the sum $s_n(x)$ only for the higher harmonics. In other words, the coefficients of the initial terms are practically the same in the Fejér sum as they are in the partial sum of the corresponding Fourier series when the total number of terms is large. This fact means that as the number of terms becomes infinite, the coefficients for any finite number of terms are actually identical. The two resulting infinite series differ only in their terms of infinite order.*

*This is admittedly a rather loose way of referring to the higher order terms in a partial sum as the number of terms in that sum is allowed to increase without limit.

From a practical point of view this circumstance may at first sight seem trivial, inasmuch as the terms of infinite order can never be reached in any term-by-term calculation. Nevertheless, there remains a significant difference in the behaviors of the two series, in that one of them exhibits

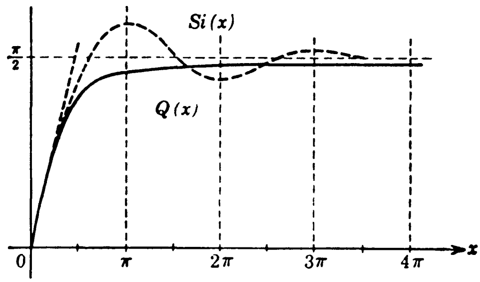


FIG. 24. Essential difference in the behavior of the partial sum of a Fourier series and that of a series of Fejér polynomials in the vicinity of a discontinuity.

the Gibbs phenomenon at any point of discontinuity and the other does not. This difference is due precisely to the difference in the terms of infinite order because these are the ones which alone are significant in determining the approximation properties of the series in the vicinity of a discontinuity.

15. Fourier analysis by graphical means

In practice it frequently occurs that the function $f(x)$ for which a Fourier series representation is wanted is available in graphical form only. Usually also in problems of this sort only a partial sum $s_n(x)$ is sought which approximates the function $f(x)$ with a certain stated tolerance. Before discussing a possible method of solution, it may be well to point out that the precise requirements of the desired solution in a problem of this kind are frequently not clearly stated, and hence that considerable confusion regarding the value of a particular solution may result.

The problem is frequently put very roughly in the statement that a finite trigonometric polynomial of the form of $s_n(x)$ is sought which approximates a given function, but usually nothing is said about the approximation properties of the desired polynomial except perhaps that the " best approximation " which a given number of terms can yield is wanted. Inasmuch as there are an infinite variety of ways in which a trigonometric polynomial with a given number of terms may approximate the required function $f(x)$, the solution is decidedly not unique. For example, the polynomial may approximate $f(x)$ so as to make the mean

and odd components $f_1(x)$ and $f_2(x)$ respectively according to the relations 80 and 81. The Fourier series for these components read

$$f_1(x) = a_0 + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + \cdots \tag{283}$$

$$f_2(x) = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots \tag{284}$$

and the corresponding partial sums are written

$$s^{(1)}{}_{n-1}(x) = a_0 + a_1 \cos x + \cdots + a_{n-1} \cos (n-1)x \tag{285}$$

$$s^{(2)}{}_n(x) = b_1 \sin x + b_2 \sin 2x + \cdots + b_n \sin nx \tag{286}$$

The problem is to determine the coefficients in these partial sums so that

$$s^{(1)}{}_{n-1}(x) \sim f_1(x) \tag{287}$$

$$s^{(2)}{}_n(x) \sim f_2(x) \tag{288}$$

According to Art. 2, the range $0 < x < \pi$ is divided into n subranges, the centers of which correspond to the x-values $0 < x_1 < x_2 < \cdots < x_n < \pi$ with

$$x_k = \frac{(2k-1)\pi}{2n} \tag{289}$$

With the coefficients

$$\alpha_{ks} = \cos sx_k \qquad \text{and} \qquad \beta_{ks} = \sin sx_k \tag{290}$$

as defined by Eqs. 52 and 53, the problem reduces to a determination of the solution of the systems of linear algebraic equations

$$\sum_{s=0}^{n-1} \alpha_{ks} a_s = f_1(x_k) \qquad (k = 1, 2, \cdots n) \tag{291}$$

and

$$\sum_{s=1}^{n} \beta_{ks} b_s = f_2(x_k) \qquad (k = 1, 2, \cdots n) \tag{292}$$

Multiplying these equations respectively by α_{kr} and β_{kr} and summing over k yields

$$\sum_{s=0}^{n-1} \left(\sum_{k=1}^{n} \alpha_{kr}\alpha_{ks} \right) a_s = \sum_{k=1}^{n} \alpha_{kr} f_1(x_k) \tag{293}$$

and

$$\sum_{s=1}^{n} \left(\sum_{k=1}^{n} \beta_{kr}\beta_{ks} \right) b_s = \sum_{k=1}^{n} \beta_{kr} f_2(x_k) \tag{294}$$

In view of the orthogonality conditions stated by Eqs. 56 and 57, one

finds

$$\sum_{k=1}^{n} \alpha_{kr} f_1(x_k) = \begin{cases} \dfrac{n}{2} a_r & \text{for } r = 1, 2, \cdots n - 1 \\ n a_0 & \text{for } r = 0 \end{cases} \quad [295]$$

and

$$\sum_{k=1}^{n} \beta_{kr} f_2(x_k) = \begin{cases} \dfrac{n}{2} b_r & \text{for } r = 1, 2, \cdots n - 1 \\ n b_n & \text{for } r = n \end{cases} \quad [296]$$

Observing that

$$\beta_{kn} = \sin n x_k = \sin (k - \tfrac{1}{2})\pi = (-1)^{k-1} \quad [297]$$

and that

$$\alpha_{k0} = 1 \quad [298]$$

Eq. 295 gives

$$a_r = \frac{2}{n} \sum_{k=1}^{n} \alpha_{kr} f_1(x_k) \qquad (r = 1, 2, \cdots n - 1) \quad [299]$$

and in particular

$$a_0 = \frac{1}{n} \sum_{k=1}^{n} f_1(x_k) \quad [300]$$

whereas Eq. 296 yields

$$b_r = \frac{2}{n} \sum_{k=1}^{n} \beta_{kr} f_2(x_k) \qquad (r = 1, 2, \cdots n - 1) \quad [301]$$

with

$$b_n = \frac{1}{n} \sum_{k=1}^{n} (-1)^{k-1} f_2(x_k) \quad [302]$$

The last four equations constitute the desired solution. The values of $f_1(x_k)$ and $f_2(x_k)$ for $k = 1, 2, \cdots n$ are taken from the graphical plots for these functions, and the values of the coefficients α_{kr} and β_{kr} are either calculated from the relations 290 or read from graphical plots of the sine and cosine functions, it being noted that the arguments of these functions are $(2k - 1)s\pi/2n$. Thus all the values of the coefficients α_{kr} and β_{kr} are readily taken from a pair of carefully plotted curves for the sine and cosine functions over a 90-degree interval. In this way the solutions given by Eqs. 299 to 302 are quite rapidly evaluated for reasonably large values of n.

It should be observed that the values of the coefficients a_r and b_r thus obtained, when substituted into the partial sums 285 and 286, yield functions which agree respectively with $f_1(x)$ and $f_2(x)$ at the points $x_1 \cdots x_n$ and, of course, also at the points which correspond to the negatives of these x-values. Over the small ranges between these points the values of the partial sums do not agree with the corresponding ones of the functions $f_1(x)$ and $f_2(x)$, but if n is chosen sufficiently large it is reasonable to expect a fairly good approximation between the partial sums and the functions which they are to represent.

If the functions $f_1(x)$ and $f_2(x)$ are relatively smooth, the points $x_1 \cdots x_n$ may be spaced farther apart; if the given functions are very irregular, the spacing of these points must be smaller. At all events the spacing must be sufficiently close to take account of the most rapidly varying portions of the given functions; otherwise such variations cannot be expected to be even approximately reproduced by the resulting partial sums.

Of the nature of the resulting approximation not much can be said inasmuch as this depends largely upon the particular characteristics of the given function and the chosen spacing of the points $x_1 \cdots x_n$. In order to conserve computational labor, this spacing is ordinarily chosen as large as possible, in which case the coefficients in the vicinity of the terms of highest order cannot be expected to be even approximately equal to the true Fourier coefficients. However, in view of the discussion given earlier in this article, it does not necessarily follow that these coefficients are practically of less value than the others.

In various practical problems, methods of determining finite trigonometric polynomials which exhibit controllable approximation properties would be highly desirable. Of particular interest in this respect would be a method of determining a polynomial which approximates the given function with a uniform tolerance, since this type yields the closest approximation for a given finite number of terms. Such methods are desirable not only for use with graphically given functions but also for analytically given ones as well. Unfortunately, these questions have as yet received little attention.

16. Relation to the Bessel functions; Sommerfeld's *integral*

The problem of determining the spectra of frequency or phase-modulated sinusoidal time functions, which occurs in the consideration of communication signals for which this type of modulation is used, reduces to the determination of the Fourier series representation for the functions $\cos(\rho \sin x)$ and $\sin(\rho \sin x)$, in which ρ is a parameter. Since these are

the real and imaginary parts of the function

$$f(x) = e^{j\rho \sin x} = \cos(\rho \sin x) + j \sin(\rho \sin x) \qquad [303]$$

one may obtain both the desired Fourier series by considering this complex function alone. Incidentally, it is to be observed that this is a complex function of the real variable x and not a function of a complex variable. No extension of the Fourier series to functions of a complex variable is involved here, although such an extension is possible.

The Fourier series for the function 303 is advantageously written in the complex form

$$f(x) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jnx} \qquad [304]$$

in which

$$\alpha_n = \frac{1}{2\pi}\int_0^{2\pi} f(x) e^{-jnx}\,dx = \frac{1}{2\pi}\int_0^{2\pi} e^{j(\rho \sin x - nx)}\,dx \qquad [305]$$

This integral is a special form of the more general complex integral

$$Z_p(\rho) = \frac{1}{\pi}\int e^{j\rho\cos\zeta} e^{jp(\zeta - \pi/2)}\,d\zeta \qquad [306]$$

which Sommerfeld has shown* to be capable of representing cylinder functions of all kinds according to the specific choice made for the path of integration in the complex ζ-plane.

For the present problem it is sufficient to observe that the form which the integral 306 takes for the representation of cylinder functions of the first kind (Bessel functions), with the order p equal to an integer n, is (except for a factor $1/2$) identical with the integral 305. This function is usually denoted by $J_n(\rho)$, thus,

$$\alpha_n = J_n(\rho) = \frac{1}{2\pi}\int_0^{2\pi} e^{j(\rho \sin x - nx)}\,dx \qquad [307]$$

The restriction, that this integral representation for the Bessel function is valid only for integer values of n, is not violated in the problem of Fourier expansion considered here.

Numerous variations in the form of the integral 307 are possible, a few of which may be worth mentioning. For example, making the change of variable $x \to x + \pi/2$ and noting that the limits on the integral may, because of the periodicity of the integrand, be changed arbitrarily as long

**Math. Ann.*, *47* (1896), 335. Also in Riemann-Weber, *Differentialgleichungen der Physik*, Vol. II, p. 454 (Vieweg, 1927).

most engineering purposes by the expressions

$$J_0(x) \cong \frac{\cos\left(x - \frac{\pi}{4}\right)}{\sqrt{\frac{\pi}{2}x}} \qquad [322]$$

and

$$J_1(x) \cong \frac{\sin\left(x - \frac{\pi}{4}\right)}{\sqrt{\frac{\pi}{2}x}} \qquad [323]$$

for values of x larger than about 4. In connection with the frequency or phase modulation problem the arguments of the Bessel functions are for

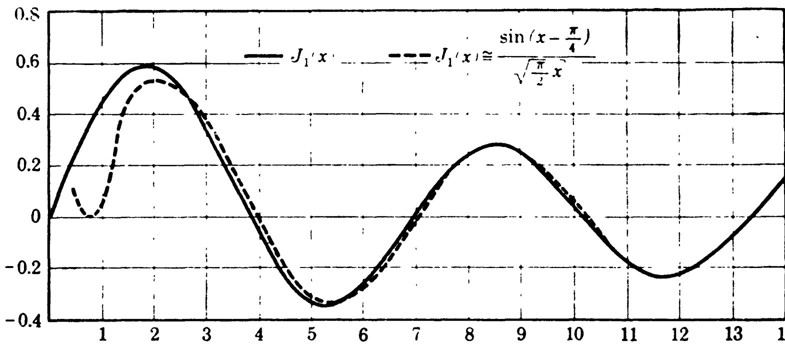


FIG. 26. A plot showing the closeness of approximation to the Bessel function of order one by the function of Eq. 323.

the most part at least as large as this value or much larger, so that it is seldom necessary to consult the tables or curves for these functions. The functions of higher order are readily calculated in terms of $J_0(x)$ and $J_1(x)$ from the recursion formula*

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x) \quad . \qquad [324]$$

*The derivation and discussion of this formula may be found in various well-known books on the subject of Bessel functions. The approximate expressions 322 and 323 are derived in Art. 26.

17. Fourier series in terms of more than one variable

The Fourier series representation of a periodic function may readily be extended to functions of more variables. For example, if $f(x,y)$ is periodic in both the variables x and y with the fundamental ranges $0 \leqq x \leqq 2\pi$ and $0 \leqq y \leqq 2\pi$, and if, throughout the rectangular region thus defined, the derivative

$$\frac{\partial^2 f}{\partial x\, \partial y} \tag{325}$$

is finite and continuous, the series

$$f(x,y) = \sum_{\mu,\nu=-\infty}^{\infty} \alpha_{\mu\nu} e^{j(\mu x+\nu y)} \tag{326}$$

is absolutely and uniformly convergent throughout the stated region and there represents the function $f(x,y)$ in the Fourier sense.

As in the case of the Fourier series for a function of a single variable, the derivative 325 or the function $f(x,y)$ need not be continuous at all points of the region. The representation 326 is still possible as long as the Dirichlet conditions (see latter part of Art. 7) are satisfied in both the variables, but the series is then no longer uniformly convergent throughout the entire fundamental region.

The coefficients in the complex series 326 are given by the formula

$$\alpha_{\mu\nu} = \frac{1}{2\pi}\int_0^{2\pi} dx\, \frac{1}{2\pi}\int_0^{2\pi} dy\, f(x,y) e^{-j(\mu x+\nu y)} \tag{327}$$

which is a straightforward extension of the formula applying to functions of a single variable. Derivation of it may be assumed to proceed by considering the function $f(x,y)$, for the moment, for a particular value of y. That is, for y = constant, $f(x,y)$ is a function o x only, and a Fourier series representation is possible in the form

$$f(x,y) = \sum_{\mu=-\infty}^{\infty} A_\mu(y) e^{j\mu x} \tag{328}$$

in which the coefficients

$$A_\mu(y) = \frac{1}{2\pi}\int_0^{2\pi} f(x,y) e^{-j\mu x}\, dx \tag{329}$$

are functions of the parameter y.

For any integer value of μ, $A_\mu(y)$ may be represented by a Fourier series in the variable y, thus

$$A_\mu(y) = \sum_{\nu=-\infty}^{\infty} \alpha_{\mu\nu} e^{j\nu y} \tag{330}$$

with coefficients

$$\alpha_{\mu\nu} = \frac{1}{2\pi}\int_0^{2\pi} A_\mu(y)e^{-j\nu y}\,dy \qquad [331]$$

Substituting Eq. 330 into Eq. 328 yields the double sum 326, which is the desired expansion for $f(x,y)$; and substituting Eq. 329 into Eq. 331 yields the formula 327, for the resulting Fourier coefficients.

Thus the Fourier expansion may readily be extended to functions of any number of variables.

18. Frequency groups

In the present discussion a finite group of simple harmonic time functions (frequency components) is assumed to be given, and an inquiry is directed toward determining the nature of the function defined by linear superposition of them. Although this problem has much in common with the consideration of the properties of the partial Fourier and Fejér sums discussed in Arts. 13 and 14, the present interpretations are directed toward an entirely different goal, namely, toward a means for representing, in a closed form, functions which are not necessarily periodic yet for which the defining range extends over the entire region of the independent variable from minus to plus infinity.

To begin with a simple example, it is assumed that three simple harmonic functions with any finite amplitudes are given, having the frequencies 100, 125, and 150 cycles per second. The relative phase angles are for the moment immaterial. The resultant function given by the sum of these three components is periodic, its fundamental frequency being 25 cycles per second. This conclusion is clear from the fact that 25 is the highest common factor (HCF) of the group of numbers 100, 125, and 150. The fundamental period is one-twenty-fifth of a second. Throughout this interval the 100 cycles per second component completes 4 cycles, the 125 cycles per second component completes 5 cycles, and the 150 cycles per second component completes 6 cycles. The original state of affairs is then re-established because each component has completed a whole number of cycles. This statement is true no matter what the relative phase angles of the components may be.

In the language of the Fourier series, the resultant function is represented by its fourth, fifth, and sixth harmonic components alone. All other components, including the fundamental, are absent. It becomes clear that the linear superposition of a group of simple harmonic components yields a periodic function only if their frequencies have a common measure. This means that the frequencies must be given by rational numbers or be rational multiples of the same irrational or transcendental

number. If several different irrational numbers (like $\sqrt{2}$) or transcendental numbers (like π or the naperian base e) are contained in the group of frequencies, the resultant function never exactly repeats its sequence of values; that is, its period is infinite. The same is true if rational, irrational, and transcendental numbers in any combination are contained in the group of frequencies.

For example, the function

$$f(t) = \cos 100t + \cos (100 + \pi)t \qquad [332]$$

never repeats its sequence of values. This situation should not be confused with the well-known circumstance that the function given by Eq. 332 can be interpreted as a beat phenomenon through conversion of the right-hand side of this equation to the form

$$f(t) = 2 \cos \frac{\pi}{2} t \cdot \cos \left(100 + \frac{\pi}{2}\right) t \qquad [333]$$

which is customarily plotted through considering the slowly varying function $2 \cos (\pi/2)t$ as an envelope containing the rapidly varying simple harmonic function $\cos (100 + \pi/2)t$. Each half period of the function $\cos (\pi/2)t$ is commonly referred to as the "beat period," but this is a true period only if the difference between the two frequencies involved in the expression 332 is at the same time their highest common factor.

Even when the two frequencies are rational numbers the beat period is not necessarily a true period. For example, the frequencies 100 cycles per second and 103 cycles per second give rise to a beat frequency of 3 cycles per second, but the true fundamental frequency is 1 cycle per second. In other words, the exact pattern of the resultant function does not repeat until three beat periods have elapsed. The well-known experimental fact that the human ear (under proper circumstances) appears to recognize the beat frequency as though it were actually present in the form of a separate component is a physiological phenomenon due in part to a nonlinear characteristic in the response mechanism of the ear, and this is not to be confused with the strictly linear superposition of component frequencies considered here.

It is interesting as well as instructive to generalize the problem of the simple beat phenomenon by inquiring how the interference pattern looks when more than two frequency components are superimposed. Incidentally, one must recognize that the familiar beat pattern in the case of two frequencies is pronounced only when the increment between these frequencies is small compared with either one. When the two frequencies and their difference are of the same order of magnitude, the conversion from the form of Eq. 332 to that of Eq. 333 is, of course, still valid, but the two functions whose product is represented by Eq.

333 then vary at about the same rate and the beat character of the resultant function is lost.

In generalizing the problem of the beat phenomenon it is, therefore, essential to assume a group of simple harmonic components whose spacing in the frequency spectrum is small compared with the mean frequency of the group. The line spectrum of such a group containing seven components is illustrated in Fig. 27. The mean angular frequency is ω_0. The adjacent frequencies are $\omega_0 + \delta\omega$, $\omega_0 - \delta\omega$, $\omega_0 + 2\delta\omega$, $\omega_0 - 2\delta\omega$, $\cdots$ and so forth, their uniform spacing being equal to $\delta\omega$.

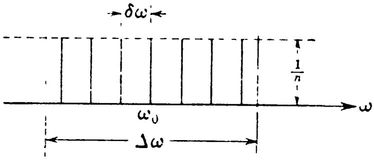


Fig. 27. The line spectrum of a frequency group.

In general, a frequency group of this sort is considered to consist of n components, and for simplicity all the amplitudes are assumed to be equal and all the phase angles zero. It is further expedient to set the common amplitude equal to $1/n$. The width of the *group* is defined as

$$\Delta\omega = n\delta\omega \tag{334}$$

The analytic expression for the group then reads

$$f(t) = \frac{1}{n}\cdot\left\{\begin{array}{l} \cos\omega_0 t + \cos(\omega_0 + \delta\omega)t + \cos(\omega_0 + 2\delta\omega)t \\ \qquad + \cdots + \cos\left(\omega_0 + \dfrac{n-1}{2}\delta\omega\right)t \\ + \cos(\omega_0 - \delta\omega)t + \cos(\omega_0 - 2\delta\omega)t \\ \qquad + \cdots + \cos\left(\omega_0 - \dfrac{n-1}{2}\delta\omega\right)t \end{array}\right\} \tag{335}$$

By repeated use of the trigonometric identity

$$\cos(a \pm b) = \cos a \cos b \mp \sin a \sin b \tag{336}$$

this expression may be put into the form

$$f(t) = \frac{2\cos\omega_0 t}{n}\cdot\left\{\frac{1}{2} + \cos\delta\omega t + \cos 2\delta\omega t + \cdots + \cos\frac{n-1}{2}\delta\omega t\right\} \tag{337}$$

With the formula expressed by Eq. 11 this result may be written

$$f(t) = \frac{\sin\dfrac{n\delta\omega}{2}t}{n\sin\dfrac{\delta\omega}{2}t}\cdot\cos\omega_0 t \tag{338}$$

Here the function

$$F(t) = \frac{\sin \frac{n\delta\omega}{2} t}{n \sin \frac{\delta\omega}{2} t} \qquad [339]$$

is slowly variable, compared to the mean frequency component $\cos \omega_0 t$, and may be regarded as an envelope function enclosing this mean frequency. The beat phenomenon is placed in evidence by the envelope function just as it is in the simple case of two interfering components.

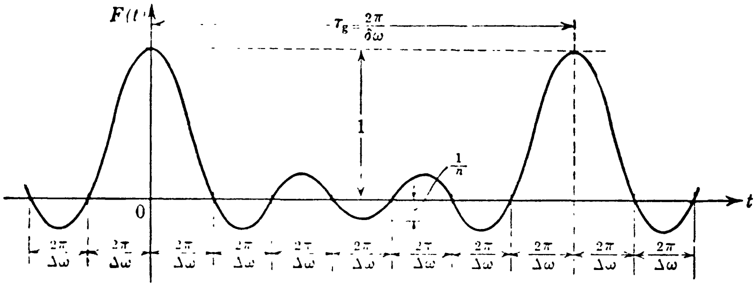


FIG. 28. The interference pattern of the frequency group given in Fig. 27.

This envelope function is plotted in Fig. 28 versus the time t for the case $n = 7$. The interference pattern for the group of frequencies whose line spectrum is given in Fig. 27 is thus illustrated by Fig. 28. The regions of constructive interference, which lie in the vicinities of the main humps of the envelope, are spaced at intervals of

$$\tau_g = \frac{2\pi}{\delta\omega} \text{ seconds} \qquad [340]$$

The duration of each main hump is $4\pi/\Delta\omega$ seconds. The group period or beat period τ_g is, according to Eq. 340, inversely proportional to the frequency increment $\delta\omega$, and the duration of a region of constructive interference is inversely proportional to the width $\Delta\omega$ of the group. If this width is kept constant while more components are added to the group, the regions of constructive interference occur at longer intervals but their duration remains the same. The number of smaller humps between the large ones also increases; but, for a large number of component frequencies, the amplitudes of these smaller humps become insignificant midway between the large humps and in this vicinity.

The fact that the interval between regions of constructive interference is inversely proportional to the frequency increment $\delta\omega$ leads to the con-

clusion that, as this increment is allowed to approach zero, the beat period grows without limit. The frequency group finally becomes continuous, and the resulting function has only one region of constructive interference in the entire time scale from minus to plus infinity. This reasoning assumes, of course, that each frequency component in the group endures throughout the entire time scale; that is, it represents a true steady state component.

The limiting form for the envelope function $F(t)$, resulting from letting $\delta\omega$ approach zero, is readily evaluated. Since the width $\Delta\omega$ of the group remains constant, n and $\delta\omega$ must vary inversely as the limit $n \rightarrow \infty$, $\delta\omega \rightarrow 0$ is carried to completion. The trigonometric sine in the denominator of Eq. 339 may be replaced by its argument, so that the limiting form of this function is readily seen to be

$$F(t) = \frac{\sin \frac{\Delta\omega}{2} t}{\frac{\Delta\omega}{2} t} \qquad [341]$$

This is the function discussed in Art 13 and plotted in Fig. 17, except that here

$$\eta = \frac{\Delta\omega}{2} t \qquad [342]$$

The resultant group function

$$f(t) = \left(\frac{\sin \frac{\Delta\omega}{2} t}{\frac{\Delta\omega}{2} t} \right) \cos \omega_0 t \qquad [343]$$

is now entirely transient in nature: that is, it never repeats, but has only one region of constructive interference. Beyond this region the components in the group (which are now infinite in number) interfere destructively forever.

Since the phase angles of all the components are chosen equal to zero, the region of constructive interference lies at the time origin. If in the argument of each cosine term in Eq. 335 a phase angle $-\phi$ is inserted which has the following frequency dependence

$$\phi(\omega) = (\omega - \omega_0)t_0 \qquad [344]$$

the variable t in the above formulas for the envelope function is replaced by $(t - t_0)$. The region of constructive interference then occurs in the vicinity of the point $t = t_0$. If the phase angles of the components in the group are given random values, the form of the resulting interference

pattern becomes very difficult to determine and in general exhibits no well-defined region of constructive interference.

The fact that a continuous group of frequency components gives rise to a resultant function having a transient nature suggests that if a similar limiting process is carried out with the Fourier series, a means will be obtained for analytically representing an arbitrary nonperiodic function. This suggestion is followed in the following article.

19. The Fourier integral

The essence of the discussion in the preceding articie may be summarized thus: Whereas the linear superposition of a group of discrete, uniformly spaced, frequency components (finite or infinite in number) gives rise to an interference pattern of a periodic nature, this pattern assumes a transient character when the frequencies in the group are continuously distributed. The resulting so-called *continuous spectrum* should be thought of as a line spectrum in which the spacing of the lines is allowed to approach zero, and the transient character of the resulting time function may be thought of as the limiting form of a periodic function for which the period has become infinite. That these two limiting processes are consistent is evident from the fact that the spacing of the lines in the spectrum of a periodic function equals its fundamental frequency, which is the inverse of its period. It may further be helpful in this connection to recognize that the period (which is the reciprocal of the spacing of the lines) is equal to the line density expressed as the number of lines per cycles per second. As the period is allowed to grow without limit, the line density grows without limit, so that finally the spectrum becomes a continuous one and the function never repeats; that is, it becomes a transient function.

By the carrying out of this limiting process in terms of the Fourier series Eq. 175 and the relation 176 for its coefficients (spectrum function), an analytic means is obtained for the representation, in a closed form, of an arbitrary transient function. The value of such a mathematical tool in connection with various engineering problems is readily appreciated.

Repeating, for convenience, the mathematical statement for the periodic case*

$$f(t) = \sum_{n=-\infty}^{\infty} \alpha_n e^{jn\omega_1 t} \qquad [345]$$

$$\alpha_n = \frac{\omega_1}{2\pi}\int_{-\pi/\omega_1}^{\pi/\omega_1} f(t)e^{-jn\omega_1 t}\,dt \qquad [346]$$

*The fundamental angular frequency is in the present article denoted by ω_1 in order that the symbol ω may be used to denote any arbitrary angular frequency.

one should observe that the limiting process indicated by $\tau \to \infty$ or $\omega_1 \to 0$ evidently is accompanied by $\alpha_n \to 0$. That is, as the line spacing in the spectrum is allowed to become smaller and smaller, the amplitudes of the harmonic components also become smaller and smaller. However, it may be expected that the ratio α_n/ω_1 approaches a finite limiting function. Thus, before carrying out the limiting process, it is expedient to rewrite Eqs. 345 and 346 in the form

$$f(t) = \sum_{n\omega_1=-\infty}^{\infty} \left(\frac{\alpha_n}{\omega_1}\right) e^{jn\omega_1 t}\,\delta(n\omega_1) \tag{347}$$

$$\left(\frac{\alpha_n}{\omega_1}\right) = \frac{1}{2\pi}\int_{-\pi/\omega_1}^{\pi/\omega_1} f(t)e^{-jn\omega_1 t}\,dt \tag{348}$$

In Eq. 347 the symbol $\delta(n\omega_1)$ stands for the increment in the variable $n\omega_1$, which is equal to the line spacing and hence equal to ω_1.

The limiting process is now formally indicated by

$$\begin{aligned} &\tau \to \infty \\ &\left.\begin{aligned}\omega_1 &\to 0 \\ n &\to \infty\end{aligned}\right\} (n\omega_1) \to \omega \\ &\delta(n\omega) \to d\omega \\ &\left(\frac{\alpha_n}{\omega_1}\right) \to g(\omega) \end{aligned} \tag{349}$$

Here it should be recognized that a new variable for the frequency, a continuous variable, must be introduced to take the place of the discontinuous variable $n\omega_1$. This new variable (denoted by ω) refers to any finite angular frequency in the continuous spectrum just as the discontinuous variable $n\omega_1$ does in the line spectrum. The introduction of the new symbol may be regarded as a convenient way of avoiding the appearance of the quantities n and ω_1, which become improper in the limit. However, $\omega = n\omega_1$ remains a proper variable and still refers to the frequency of any harmonic component in the limit precisely as it does before this limit is carried to completion.

If this limiting process is thought of as being carried out in steps through doubling and redoubling of the period τ, it becomes clear that each time τ is doubled, any specific harmonic component doubles its order n. If attention is fixed upon a specific harmonic frequency, n and ω_1 must vary inversely (that is, $n\omega_1$ must remain constant) as the period is increased. The frequency increment $\delta(n\omega_1)$ in the limit is formally replaced by the differential $d\omega$, and the ratio α_n/ω_1 becomes a finite function $g(\omega)$ of the continuous frequency variable ω. This function expresses the variation of the harmonic amplitudes in the limit.

It remains to recognize that as the limiting process is carried to com-

pletion, the summation in Eq. 347 becomes an integration. The final forms of the relations 347 and 348 are

$$f(t) = \int_{-\infty}^{\infty} g(\omega)\, d\omega\, e^{j\omega t} \tag{350}$$

$$g(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, dt\, e^{-j\omega t} \tag{351}$$

This heuristic derivation of the relations 350 and 351 does not establish their correctness on a rigorous mathematical basis, but from an engineering point of view a rigorous proof may properly be omitted since the principal interest lies in the interpretation and use of these forms.

It may be pointed out once more that the complex harmonic amplitudes are not given by $g(\omega)$ but by $g(\omega)\, d\omega$. Since $g(\omega)$ is finite, and $d\omega$ is the symbolic representation of a quantity which is regarded as becoming vanishingly small, the harmonic amplitudes are also vanishingly small. However, just as $\delta(n\omega_1)$ denotes a constant spacing in the line spectrum of a periodic function, so $d\omega$, which represents $\delta(n\omega_1)$ as the limiting process is carried to completion, must, at any stage of this process, likewise be regarded as a constant. Hence $g(\omega)$ is proportional to $g(\omega)\, d\omega$, so that a plot of $g(\omega)$ versus ω shows the correct variation of the harmonic amplitudes with frequency even though all these amplitudes are vanishingly small.

Equation 350 is called the *Fourier integral* representation for the function $f(t)$. The function $g(\omega)$ is called the *Fourier transform* of $f(t)$ and, reciprocally, $f(t)$ is called the *inverse* Fourier transform of $g(\omega)$. The second of the pair of integrals 350 and 351 transforms a time function $f(t)$ into its equivalent frequency function $g(\omega)$, and the first of these integrals reverses the process. The second integral analyzes the time function into a spectrum, and the first integral synthesizes the spectrum to regain the time function. $g(\omega)$ represents the function in the frequency domain just as $f(t)$ represents the function in the time domain. One may also regard Eqs. 350 and 351 as representing simultaneously a pair of integral equations and their mutual solutions.

The graphical interpretation of the entire process of Fourier analysis and synthesis in terms of the frequency spectrum as discussed in Art. 9, as well as the method of manipulation of the forms 350 and 351 in connection with various physical and mathematical problems, remains exactly the same as for the Fourier series. Hence there is nothing new to be learned in this respect. The essential difference lies only in the fact that the spectrum function is continuous and that the synthesis of the spectrum is accomplished by means of an integral instead of a sum. This latter difference is actually an advantage because more formulas are

available for the evaluation of integrals than for the evaluation of sums. The principal advantage, however, lies in the fact that the Fourier integral is capable of representing transient functions.

The conditions under which this representation is possible are essentially the same as those which apply to the representation of a periodic function by means of a Fourier series. These are the Dirichlet conditions as pointed out in Art. 7. A detailed difference in the form of these conditions arises from the fact that the fundamental range is now infinite instead of being finite. In view of this difference, the condition 136 reads

$$\int_{-\infty}^{\infty} |f(t)|\, dt \text{ shall be finite} \tag{352}$$

At a discontinuity of the function $f(t)$ the Fourier integral (like the Fourier series) also yields the arithmetic mean between the two immediately adjacent values of the function (as stated for the Fourier series by the relation 137).

The approximation properties of the Fourier integral are likewise the same as those of the series. In order to show this, one may consider the integral 350 for finite limits. Thus, the function

$$s(t) = \int_{-a}^{a} g(\omega)\, d\omega\, \epsilon^{j\omega t} \tag{353}$$

is the analogue of the partial sum of a Fourier series, since it represents the synthesis of a finite portion of the spectrum. Substituting for $g(\omega)$ the expression given by Eq. 351, and using θ in place of the variable t, one has after interchanging the order of the two integrations

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(\theta)\, d\theta \int_{-a}^{a} \epsilon^{-j\omega(\theta-t)}\, d\omega \tag{354}$$

Writing for the exponential

$$\epsilon^{-j\omega(\theta-t)} = \cos\omega(\theta - t) - j\sin\omega(\theta - t) \tag{355}$$

and observing that the cosine is even while the sine is odd, one finds that

$$\int_{-a}^{a} \epsilon^{-j\omega(\theta-t)}\, d\omega = 2\int_{0}^{a} \cos\omega(\theta - t)\, d\omega \tag{356}$$

and hence that Eq. 354 may be written

$$s(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(\theta)\, d\theta \int_{0}^{a} \cos\omega(\theta - t)\, d\omega \tag{357}$$

But

$$\int_{0}^{a} \cos\omega(\theta - t)\, d\omega = \frac{\sin a(\theta - t)}{(\theta - t)} \tag{358}$$

so that

$$s(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(\theta)\cdot\frac{\sin a(\theta - t)}{(\theta - t)}\cdot d\theta \qquad [359]$$

which is analogous to the integral 245 for the partial sum of a Fourier series.

In order to illustrate the application of this formula one may consider the function defined by

$$\begin{aligned} f(\theta) &= 0 \quad \text{for} \quad \theta < -\frac{\delta}{2} \\ f(\theta) &= 1 \quad \text{for} \quad -\frac{\delta}{2} < \theta < \frac{\delta}{2} \\ f(\theta) &= 0 \quad \text{for} \quad \theta > \frac{\delta}{2} \end{aligned} \qquad [360]$$

Then the integral 359 yields

$$s(t) = \frac{1}{\pi}\int_{-\delta/2}^{\delta/2} \frac{\sin a(\theta - t)}{(\theta - t)}\, d\theta \qquad [361]$$

which may be written

$$s(t) = \frac{1}{\pi}\int_{-a(\delta/2+t)}^{a(\delta/2-t)} \frac{\sin u}{u}\, du \qquad [362]$$

or

$$s(t) = \frac{1}{\pi}\int_{0}^{a(\delta/2-t)} \frac{\sin u}{u}\, du - \frac{1}{\pi}\int_{0}^{-a(\delta/2+t)} \frac{\sin u}{u}\, du \qquad [363]$$

Utilizing the definition of the *Si*-function according to Eq. 257, and observing that $Si\,(-x) = -Si\,(x)$, one obtains the result

$$s(t) = \frac{1}{\pi}\left\{Si\, a\left(t + \frac{\delta}{2}\right) - Si\, a\left(t - \frac{\delta}{2}\right)\right\} \qquad [364]$$

Figure 29 shows a plot of this result for the choice $a = 16\pi/\delta$. The separate *Si*-functions are also drawn (dotted) in order to illustrate more clearly how the resultant function is obtained. The similarity is very evident between this result and that shown in Fig. 21 for the partial sum of a Fourier series representing the periodic repetition of the present function $f(\theta)$. In fact, the approximate evaluation of the Fourier sum $s_n(x)$ for large values of n given in Art. 13 in connection with the problem illustrated in Fig. 21 likewise leads to an expression in terms of the *Si*-function.

The characteristic 18 per cent overswing at the points of discontinuity, yielding in the limit $a \to \infty$ the Gibbs phenomenon, is again in evidence, just as it is for the Fourier series. Equation 364 and the corresponding plot in Fig. 29 also illustrate the fact that the value of the Fourier integral at the points $t = \pm\delta/2$ equals the arithmetic mean between the two values of the function immediately adjacent to these points of discontinuity.

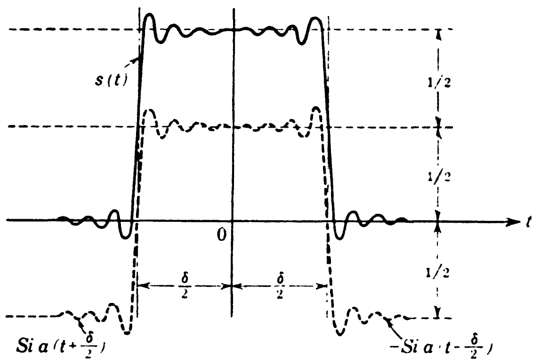


FIG. 29. The Gibbs phenomenon and the arithmetic mean property are characteristic of the Fourier integral as well as of the series.

20. ALTERNATIVE FORMS IN WHICH THE FOURIER INTEGRALS MAY BE WRITTEN

The pair of mutually inverse integral relations 350 and 351 may be written in a variety of different forms, with some of which it is well to be acquainted. One of the astonishing features about these relations is their almost complete identity in form. Thus, the integral 351, except for an interchange of the variables, ω and t, differs from the integral 350 only in the appearance of the factor $1/2\pi$ and the reversed algebraic sign in the exponent of e. The first of these differences may be removed, if removal seems desirable, by redefinition of the transform of $f(t)$ as

$$g^*(\omega) = \sqrt{2\pi}g(\omega) \tag{365}$$

The relations 350 and 351 then assume the more symmetrical forms

$$f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g^*(\omega)\, d\omega\, e^{j\omega t} \tag{366}$$

$$g^*(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\, dt\, e^{-j\omega t} \tag{367}$$

Alternatively, the appearance of a factor before these integrals may be entirely avoided through considering the transform to be a function of the cyclic frequency f (in cycles per second) instead of the angular frequency $\omega = 2\pi f$. Inasmuch as $d\omega = 2\pi\, df$, it is readily seen that when

$$\hat{g}(f) = 2\pi g(\omega) \qquad [368]$$

the integrals 350 and 351 become respectively*

$$h(t) = \int_{-\infty}^{\infty} \hat{g}(f)\, df\, e^{j2\pi ft} \qquad [369]$$

and

$$\hat{g}(f) = \int_{-\infty}^{\infty} h(t)\, dt\, e^{-j2\pi ft} \qquad [370]$$

21. Special forms for the Fourier integrals when the given function is even or odd

If the Fourier integrals 350 and 351 are written in the more explicit form

$$f(t) = \int_{-\infty}^{\infty} g(\omega)\{\cos \omega t + j \sin \omega t\}\, d\omega \qquad [371]$$

and

$$g(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\{\cos \omega t - j \sin \omega t\}\, dt \qquad [372]$$

then, if $f(t)$ is an even function of t, the second term in the integrand of 372 contributes nothing to the value of this integral, and if $f(t)$ is an odd function of t, the first term contributes nothing. In the first of these cases, moreover, $g(\omega)$ must be an even function of ω since the cosine is an even function, and in the second case, $g(\omega)$ is an odd function because the variable ω is then contained only in the argument of the sine. Hence it follows that if $f(t)$ is even, $g(\omega)$ is also even, and the Fourier integrals read

$$f(t) = \int_{-\infty}^{\infty} g(\omega) \cos \omega t\, d\omega \qquad [373]$$

$$g(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t) \cos \omega t\, dt \qquad [374]$$

*A different symbol is used here to denote the time function in order to avoid confusion with the symbol f used for the cyclic frequency.

If $f(t)$ is odd, $g(\omega)$ is also odd, and the Fourier integrals have the form

$$f(t) = j\int_{-\infty}^{\infty} g(\omega)\sin\omega t\,d\omega \qquad [375]$$

$$g(\omega) = \frac{1}{2\pi j}\int_{-\infty}^{\infty} f(t)\sin\omega t\,dt \qquad [376]$$

In both these cases the integration may alternately be extended over the range 0 to ∞ and the integrals multiplied by 2. This factor of 2 may, if desired, be absorbed by the function $g(\omega)$ in the integrals 373 and 375, whence the factors for the integrals 374 and 376 become $1/\pi$ instead of $1/2\pi$. The factor j appearing in the forms 375 and 376 evidently need not appear if the transform of $f(t)$ for this case is defined as $jg(\omega)$ and is denoted by a single symbol.

These results show that if the given function $f(t)$ is real (although this is usually the case in practice, the validity of the Fourier integrals does not require it), $g(\omega)$ is real when $f(t)$ is even, and purely imaginary when $f(t)$ is odd. In general, for real functions $f(t)$, one may decompose this function into its even and odd components $f_1(t)$ and $f_2(t)$ respectively, according to the relations 80 and 81, and then have for the corresponding transform

$$g(\omega) = g_1(\omega) + jg_2(\omega) \qquad [377]$$

in which

$$g_1(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(t)\,dt\,e^{-j\omega t} = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(t)\cos\omega t\,dt \qquad [378]$$

and

$$g_2(\omega) = \frac{1}{2\pi j}\int_{-\infty}^{\infty} f_2(t)\,dt\,e^{-j\omega t} = -\frac{1}{2\pi}\int_{-\infty}^{\infty} f_2(t)\sin\omega t\,dt \qquad [379]$$

22. Some elementary properties of the Fourier transforms

In the Fourier integrals 350 and 351, if the algebraic signs of both variables ω and t are reversed (ω replaced by $-\omega$ and t by $-t$), the forms of these integrals remain unchanged except for the appearance of $f(-t)$ in place of $f(t)$ and $g(-\omega)$ in place of $g(\omega)$. To see this, one should note first that changing the signs of both ω and t leaves the exponents of e unchanged; and, second, that the signs of $d\omega$ and dt do change but so do the signs of the infinite limits, and since this amounts to an interchange of the limits, the resulting signs of the integrals are the same as before. Hence, one obtains the result:

If g(ω) *is the transform of* f(t), *then* g($-\omega$) *is the transform of* f($-$t) [380]

In the general case in which both $f(t)$ and $g(\omega)$ may be complex, it should be observed that both sides of Eqs. 350 and 351 assume their conjugate values if the functions $f(t)$ and $g(\omega)$ are replaced by their conjugate values $\bar{f}(t)$ and $\bar{g}(\omega)$, and if, in addition, the algebraic sign of either ω or t (not of both) is changed. This change is equivalent to changing the sign of j in the exponent of e. The statement, therefore, follows that:

$$\begin{aligned}&\textit{If } g(\omega) \textit{ is the transform of } f(t), \textit{ then}\\ &\bar{g}(\mp\omega) \textit{ is the transform of } \bar{f}(\pm t)\end{aligned} \qquad [381]$$

As pointed out in Art. 20, the integral 350, which performs the inverse of the transformation given by the integral 351, is almost identical to the latter in form. If it were identical in form, one could omit entirely the distinction between the transform and the inverse transform of a given function, since two successive applications of the same transformation would yield the function itself. That is, the variables ω and t might be interchanged, or in other words $f(t)$ and $g(\omega)$ might be regarded as each other's transform.

Although the actual situation is not quite so simple as this, it may be observed that the forms of the two integrals do become interchanged if (*a*) Eq. 351 is multiplied by 2π, (*b*) the integral 350 is multiplied by $1/2\pi$ and the function $g(\omega)$ under this integral sign by 2π, and (*c*) the algebraic sign of either ω or t is reversed. One may then interchange the variables ω and t and regard g as the given function and f as its transform or equivalent frequency function. This conclusion is summarized by the statement:

$$\begin{aligned}&\textit{If } g(\omega) \textit{ is the transform of } f(t), \textit{ then}\\ &f(\pm\omega) \textit{ is the transform of } 2\pi g(\mp t)\end{aligned} \qquad [382]$$

In particular, if $f(t)$ is even so that $g(\omega)$ is even also, the process of changing the algebraic sign of either ω or t may be omitted. It is then true that if $g(\omega)$ is the transform of $f(t)$, one may reciprocally regard $2\pi g$ as a given function of time and have f represent the corresponding frequency function.

The statement 382 may be given an even simpler form if it is made in terms of the functions $f(t)$ and $g^*(\omega)$ defined by the integrals 366 and 367, or in terms of the functions $h(t)$ and $\hat{g}(f)$ defined by the integrals 369 and 370. The factor 2π then does not mar the almost complete reciprocity existing between the pair of functions involved in the statement.

Some problems make it convenient or necessary to change the time scale for the given function $f(t)$ by some factor; it is then desirable to know the effect of this change on the corresponding transform. In order to see the effect, one replaces the variable t in the integral of Eq. 351 by

at. As a result, the differential dt becomes replaced by a times dt and the exponent of e becomes $-j\omega at$. If now one replaces ω by ω/a, there results

$$\frac{1}{a}g\left(\frac{\omega}{a}\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(at)\, dt\, e^{-j\omega t} \qquad [383]$$

Hence it follows that:

If $g(\omega)$ *is the transform of* $f(t)$, *then*
$\frac{1}{a}g\left(\frac{\omega}{a}\right)$ *is the transform of* $f(at)$ [384]

This statement is true only if the factor a is positive. For example, for $a = -1$ the statement 384 is evidently incorrect because it conflicts with the statement 380. This conflict arises because the algebraic signs of the limits change when a is replaced by $-a$.

Another useful result follows from the observation that if the function $f(t)$ in the integral 351 is replaced by the product $f(t) \cdot e^{\pm j\omega_0 t}$, and the two exponentials in the resulting integrand are combined into one exponential with the exponent $-j(\omega \mp \omega_0)t$, the effect upon the function g is merely to replace its independent variable ω by $(\omega \mp \omega_0)$. Hence:

If $g(\omega)$ *is the transform of* $f(t)$, *then*
$g(\omega \mp \omega_0)$ *is the transform of* $f(t) \cdot e^{\pm j\omega_0 t}$ [385]

The latter function is the complex form of a sinusoidal function of angular frequency ω_0 whose amplitude is modulated by the function $f(t)$. The corresponding spectrum function is seen to be the same as that for $f(t)$ except for a translation equal to the value of ω_0. Utilizing this result together with the principle of linear superposition, one may readily obtain the spectrum functions corresponding to $f(t) \cdot \cos \omega_0 t$ or $f(t) \cdot \sin \omega_0 t$.

The complement to the statement 385 is obtained through assuming the function $g(\omega)$ in the integral 350 to be replaced by $g(\omega)e^{\pm j\omega t_0}$. Again combining the exponentials, one finds that:

If $g(\omega)$ *is the transform of* $f(t)$, *then*
$g(\omega) \cdot e^{\pm j\omega t_0}$ *is the transform of* $f(t \pm t_0)$ [386]

This result was observed in connection with the Fourier series in Art. 4, namely, that a displacement of the time function merely amounts to adding increments to the harmonic phase angles which are linearly proportional to the respective harmonic frequencies.

Other useful relationships between the transforms are found through observing that it is permissible to differentiate or integrate with respect to the parameter contained in the integrand of either of the integrals 350 or 351 as long as the resulting functions involved in these expressions still fulfill the conditions for Fourier integral representation. For example,

differentiating both sides of Eq. 350 with respect to t yields

$$\frac{df}{dt} = \int_{-\infty}^{\infty} j\omega g(\omega)\, d\omega\, e^{j\omega t} \tag{387}$$

Repeating the process n times, one finds

$$f^{(n)}(t) = \int_{-\infty}^{\infty} (j\omega)^n g(\omega)\, d\omega\, e^{j\omega t} \tag{388}$$

from which one may conclude that:

If $g(\omega)$ *is the transform of* $f(t)$, *then* $(j\omega)^n g(\omega)$ *is the transform of the* n*th derivative of* $f(t)$ [389]

The function given by this nth derivative of $f(t)$ must, of course, still fulfill the requirements for which its representation by means of the Fourier integral is valid. If the function given by the integral of $f(t)$, namely,

$$F(\theta) = \int f(\theta)\, d\theta \tag{390}$$

also fulfills these conditions, then since

$$\frac{d}{dt}\left[\int_{-\infty}^{t} f(\theta)\, d\theta\right] = \frac{dF}{dt} = f(t) \tag{391}$$

it may be inferred from the statement 389 that:

If $g(\omega)$ *is the transform of* $f(t)$, *then* $g(\omega)/j\omega$ *is the transform of* $\int_{-\infty}^{t} f(\theta)\, d\theta$ [392]

Under the same conditions, this statement may be extended to functions formed by successive integrations of $f(t)$.

Similarly, by differentiating or integrating Eq. 351 with respect to ω, one finds that:

If $g(\omega)$ *is the transform of* $f(t)$, *then the* n*th derivative of* $g(\omega)$ *is the transform of* $(-jt)^n f(t)$ [393]

or:

If $g(\omega)$ *is the transform of* $f(t)$, *then* $\int_{-\infty}^{\omega} g(\mu)\, d\mu$ *is the transform of* $f(t)/-jt$ [394]

Again, the functions $(-jt)^n f(t)$ and $f(t)/-jt$ must still fulfill the conditions for Fourier integral representation.*

*The conditions under which the statements 389, 392, 393, and 394 are valid are actually far less rigid than the Dirichlet conditions provided one interprets the results in the light of Art. 24.

23. The transform of a product and the interpretation of power products and effective values for transient functions

When the given function $f(t)$ is expressed as a product of two component functions as

$$f(t) = f_1(t) f_2(t) \tag{395}$$

one is interested to know how the transform may be expressed in terms of the individual transforms of the components $f_1(t)$ and $f_2(t)$.

It is helpful to observe that this problem is very similar to that of determining the expression for the coefficients in the resultant series

$$y = y_1 y_2 = \sum_{r=-\infty}^{\infty} \alpha_r z^r \tag{396}$$

in terms of the coefficients in the component series

$$y_1 = \sum_{m=-\infty}^{\infty} a_m z^m \tag{397}$$

and

$$y_2 = \sum_{n=-\infty}^{\infty} b_n z^n \tag{398}$$

Forming the product of these two series, one has

$$y = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_m b_n z^{m+n} \tag{399}$$

or, letting

$$m + n = r \tag{400}$$

one may write this as

$$y = \sum_{r=-\infty}^{\infty} \left(\sum_{m=-\infty}^{\infty} a_m b_{r-m} \right) z^r \tag{401}$$

in which the substitution of the summation index r for n is permissible since the summation over m is independent of that over n. A comparison of Eqs. 396 and 401 reveals that the desired relationship reads

$$\alpha_r = \sum_{m=-\infty}^{\infty} a_m b_{r-m} \tag{402}$$

Similarly, if

$$f_1(t) = \int_{-\infty}^{\infty} g_1(\mu)\, d\mu\, e^{j\mu t} \tag{403}$$

and

$$f_2(t) = \int_{-\infty}^{\infty} g_2(\nu)\, d\nu\, e^{j\nu t} \tag{404}$$

are the Fourier integral representations for the component time functions in Eq. 395, one may write

$$f_1(t)f_2(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(\mu)g_2(\nu)e^{j(\mu+\nu)t}\, d\mu\, d\nu \qquad [405]$$

since μ and ν are completely independent variables.

It is now possible to make the change of variable

$$\nu = \omega - \mu; \qquad d\nu = d\omega \qquad [406]$$

because μ is a constant parameter as far as the integration on ν is concerned in Eq. 405.

Equation 405 then becomes

$$f(t) = \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} g_1(\mu)g_2(\omega - \mu)\, d\mu\, e^{j\omega t} \qquad [407]$$

If this is written

$$f(t) = \int_{-\infty}^{\infty} g(\omega)\, d\omega\, e^{j\omega t} \qquad [408]$$

the transform of $f(t)$ is seen to be given by

$$g(\omega) = \int_{-\infty}^{\infty} g_1(\mu)g_2(\omega - \mu)\, d\mu \qquad [409]$$

which is the desired relation. It is entirely analogous to the result expressed by Eq. 402 for the corresponding problem in terms of series.

Since there is no need for distinguishing between the functions $f_1(t)$ and $f_2(t)$ in this argument, it is evident that Eq. 409 remains true if the subscripts 1 and 2 are interchanged.

In an exactly analogous fashion one finds that if

$$g_1(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(\theta)\, d\theta\, e^{-j\omega\theta} \qquad [410]$$

and

$$g_2(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_2(\tau)\, d\tau\, e^{-j\omega\tau} \qquad [411]$$

then the inverse transform of the function

$$g(\omega) = g_1(\omega)g_2(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, dt\, e^{-j\omega t} \qquad [412]$$

is given by*

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(\theta)f_2(t-\theta)\,d\theta \qquad [413]$$

Here, also, the subscripts 1 and 2 may be interchanged.

Expressing $f(t)$ in this last result in terms of its Fourier integral, one has

$$\int_{-\infty}^{\infty} g_1(\omega)g_2(\omega)\,d\omega\,e^{j\omega t} = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(\theta)f_2(t-\theta)\,d\theta \qquad [414]$$

Setting $t = 0$ and subsequently substituting the symbol t for θ in the integral on the right of Eq. 414 yield the result

$$\int_{-\infty}^{\infty} f_1(t)f_2(\mp t)\,dt = 2\pi\int_{-\infty}^{\infty} g_1(\omega)g_2(\pm\omega)\,d\omega \qquad [415]$$

in which the statement 380 is also used.

This relation is analogous to the result stated for the Fourier series by Eqs. 190 and 196, in Art. 10. In other words, the left-hand side of Eq. 415 may represent the integrated power product for some physical system, whence the right-hand side expresses this integrated power in terms of the transforms corresponding to the given time functions. The relation 415 is applicable to transient functions in the same way that Eqs. 190 and 196 are to periodic functions.

If, in particular,

$$f_2(\mp t) = \bar{f}_1(\mp t) \qquad [416]$$

then, according to the statement 381

$$g_2(\mp\omega) = \bar{g}_1(\pm\omega) \qquad [417]$$

or

$$g_2(\pm\omega) = \bar{g}_1(\mp\omega) \qquad [418]$$

The relation 415 for this special case reads

$$\int_{-\infty}^{\infty} |f_1(t)|^2\,dt = 2\pi\int_{-\infty}^{\infty} |g_1(\omega)|^2\,d\omega \qquad [419]$$

which for transient time functions expresses an analogous relation to that stated by Eqs. 200 and 201 for periodic functions. The square root of the value given by Eq. 419 may be called the effective value of the transient time function. According to the definition of the effective value of a periodic function, one forms the square root of the *mean* of the integrated squared values of that function over a period. The interpre-

*The mathematical procedure described by this integral is called " convolution " (in the German literature the term used is " faltung ").

tation of the relation 419 offers a complete parallelism between the periodic and the transient cases.

Finally, it is possible to form another relationship which at times becomes practically useful. In Eq. 415, $f_2(\mp t)$ is any time function, and $g_2(\pm\omega)$ is its transform with the reversed algebraic signs of its independent variable. If the time function is chosen to be $2\pi g_2(\mp t)$, then, according to the statement 382, its transform with reversed signs of its variable is $f_2(\mp\omega)$. Since the corresponding signs of the variables t and ω are now alike, the $\mp$ signs may be dropped, and one has in place of Eq. 415,

$$\int_{-\infty}^{\infty} f_1(t)g_2(t)\,dt = \int_{-\infty}^{\infty} g_1(\omega)f_2(\omega)\,d\omega \tag{420}$$

In connection with this relation it should be carefully observed that the functions g_1 and g_2 are understood to be the Fourier transforms of the functions f_1 and f_2, and the latter are the inverse transforms of g_1 and g_2. The reason for this emphasis is that the reader may make the mistake of regarding $f_2(\omega)$ as the transform of $g_2(t)$ because the latter is written as a function of t and the former as a function of ω. This confusion may best be avoided through adopting an entirely different symbol for the independent variable, and using the same symbol for both integrations since the distinction between *given function* and *transform* is expressed respectively by the symbols f and g, and has nothing to do with the symbols used for the independent variable or variable of integration. The preferable form of Eq. 420, therefore, reads

$$\int_{-\infty}^{\infty} f_1(x)g_2(x)\,dx = \int_{-\infty}^{\infty} g_1(x)f_2(x)\,dx \tag{421}$$

24. Some illustrative examples; the singularity functions

The first function to be considered is defined by

$$\begin{aligned} f(t) &= 0, &&\text{for } t < 0 \\ f(t) &= e^{-\alpha t}, &&\text{for } t > 0 \end{aligned} \tag{422}$$

This function is illustrated in Fig. 30. According to Eq. 351 its transform is given by the integral

$$g(\omega) = \frac{1}{2\pi}\int_0^{\infty} e^{-(\alpha+j\omega)t}\,dt \tag{423}$$

which yields

$$g(\omega) = \frac{1}{2\pi(\alpha + j\omega)} \tag{424}$$

This result may be written

$$g(\omega) = g_1(\omega) + jg_2(\omega) = \frac{\alpha}{2\pi(\alpha^2 + \omega^2)} - j\frac{\omega}{2\pi(\alpha^2 + \omega^2)} \quad [425]$$

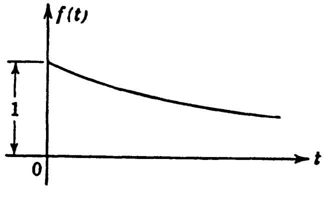


FIG. 30. An approximation to the unit step by a decaying exponential.

whence

$$g_1(\omega) = \frac{\alpha}{2\pi(\alpha^2 + \omega^2)} \quad [426]$$

and

$$g_2(\omega) = \frac{-\omega}{2\pi(\alpha^2 + \omega^2)} \quad [427]$$

According to the relations 377, 378, and 379, these are the transforms of the even and odd components respectively of the function $f(t)$. The odd components are shown plotted in parts (a) and (b) of Fig. 31. Both these components are given by $\frac{1}{2}e^{-\alpha t}$ for $t > 0$, but $f_1(t)$ is symmetrical about the vertical axis at the origin, whereas

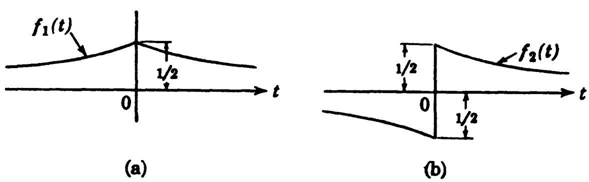


FIG. 31. Even and odd components of the function of Fig. 30.

$f_2(t)$ is antisymmetrical. The results given by Eqs. 426 and 427 may alternatively be obtained from the integrals 378 and 379 respectively. Since the integrands in these integrals are even functions, the same value is obtained if the integration is extended only over the range 0 to ∞ and the result multiplied by 2. Thus it must be true that

$$g_1(\omega) = \frac{1}{2\pi}\int_0^\infty e^{-\alpha t}\cos\omega t\,dt = \frac{\alpha}{2\pi(\alpha^2 + \omega^2)} \quad [428]$$

and

$$g_2(\omega) = \frac{-1}{2\pi}\int_0^\infty e^{-\alpha t}\sin\omega t\,dt = \frac{-\omega}{2\pi(\alpha^2 + \omega^2)} \quad [429]$$

which, incidentally, is a simple way of evaluating these two particular integrals (the usual process requires two successive integrations by parts).

These functions are shown plotted versus the ratio ω/α in Fig. 32.

Hence the point unity, for example, corresponds to $\omega = \alpha$, so that, as α is given smaller and smaller values, this point corresponds to smaller and smaller values of ω. At the same time it should be observed that the ordinates of these curves are inversely proportional to α. As α is assumed to become smaller, the function $g_1(\omega)$ appears to become more peaked, until in the limit $\alpha \to 0$ it degenerates into a single infinite ordinate at $\omega = 0$, and is zero everywhere else.

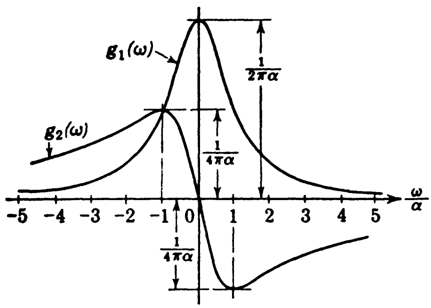


FIG. 32. The transforms of the even and odd parts given in Fig. 31.

It should be observed, however, that the area under the curve $g_1(\omega)$ is constant and independent of α, for it represents, according to the Fourier integral 350, the value of $f_1(t)$ for $t = 0$, that is,

$$f_1(0) = \int_{-\infty}^{\infty} g_1(\omega)\, d\omega = \tfrac{1}{2} \qquad [430]$$

As α becomes smaller this area becomes more and more concentrated in the immediate vicinity of the point $\omega = 0$, until in the limit $\alpha \to 0$ it is contained within a region of vanishing width at the origin. Since the area remains finite, the ordinate of $g_1(\omega)$ at $\omega = 0$ must clearly become infinite in this limit. It is quite significant that the function $g_1(\omega)$, therefore, does not vanish for $\alpha = 0$, as one might at first glance conclude by inspection of the function 426.

From Eq. 427 it is, on the other hand, easily seen that

$$\operatorname*{limit}_{\alpha \to 0} [g_2(\omega)] = \frac{-1}{2\pi\omega} \qquad [431]$$

This is the equation of a rectangular hyperbola. The manner in which this limiting function is approached by $g_2(\omega)$ as α is assumed to become smaller and smaller is readily visualized from an inspection of Fig. 32.

Turning now to the given function $f(t)$, one sees that in the limit $\alpha \to 0$ this function is zero for $t < 0$ and equal to the constant value

unity for $t > 0$. In Fig. 30, the curve for $t > 0$ then no longer falls off with increasing time but maintains the ordinate at $t = 0$ for all positive values of t. The limiting forms of the even and odd components shown in Fig. 31 are also readily visualized. $f_1(t)$ reduces to the constant $\frac{1}{2}$; $f_2(t)$ equals the constant value $-\frac{1}{2}$ for $t < 0$, equals the constant value $+\frac{1}{2}$ for $t > 0$, and retains the discontinuity of unit magnitude at $t = 0$.

This limiting function, defined by the relations 422 for $\alpha = 0$, actually does not possess a Fourier integral representation, for it no longer fulfills the condition 352. As long as α has a nonzero value, however, no matter how small, the Fourier representation is possible, and hence, for a proper interpretation of the limiting forms of the functions $g_1(\omega)$ and $g_2(\omega)$, such a representation may be said to be possible even in the limit $\alpha = 0$.

With the use of Eqs. 373 and 375 one has

$$f(t) = \int_{-\infty}^{\infty} g_1(\omega) \cos \omega t \, d\omega - \int_{-\infty}^{\infty} g_2(\omega) \sin \omega t \, d\omega \qquad [432]$$

According to the preceding discussion of the function $g_1(\omega)$, it is clear that as α becomes very small, the total contribution to the first of these two integrals is due almost entirely to the values of the integrand in the immediate vicinity of the point $\omega = 0$. If the variable t is for the moment assumed to remain finite, then for a sufficiently small value of α, the function $\cos \omega t$ remains equal to unity over that small range in the vicinity of $\omega = 0$ which contributes almost wholly to the value of this first integral. In the limit $\alpha \to 0$ this reasoning becomes exact, and in view of Eq. 430 one may, therefore, conclude that in or near this limit Eq. 432 is equivalent to

$$f(t) = \frac{1}{2} - \int_{-\infty}^{\infty} g_2(\omega) \sin \omega t \, d\omega \qquad [433]$$

Substituting the limiting value of the function $g_2(\omega)$, as expressed by Eq. 431, one has

$$f(t) = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin \omega t}{\omega} d\omega \qquad [434]$$

which may be written

$$f(t) = \frac{1}{2} + \operatorname*{limit}_{a \to \infty} \left\{ \frac{1}{2\pi} \int_{-at}^{at} \frac{\sin u}{u} du \right\} \qquad [435]$$

According to the discussion of the *Si*-function in Art. 13, and with the help of Eq. 257 the last result may be written,

$$f(t) = \frac{1}{2} + \operatorname*{limit}_{a \to \infty} \left\{ \frac{1}{\pi} Si\,(at) \right\} \qquad [436]$$

Hence it is clear that, except for the Gibbs phenomenon at the point of discontinuity, the Fourier integral representation for this function is valid even in the limit $\alpha = 0$, provided the integrals are properly interpreted.

It should be observed that one cannot simply set α equal to zero in the expression 424 for the transform of $f(t)$, for then $g(\omega)$ becomes identical with $jg_2(\omega)$, and $g_1(\omega)$, which contributes the value $\frac{1}{2}$ to the Fourier integral representation, is lost entirely. As long as α is retained as a small quantity, and discarded only after the proper interpretation of the various steps in the process of evaluation, no difficulties are encountered.

In order to illustrate this point from a slightly different angle, one may substitute the value 424 for $g(\omega)$ into the Fourier integral 350 and have

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{j\omega t}}{\alpha + j\omega}\, d\omega \qquad [437]$$

With the sine and cosine equivalent of the exponential function, this is

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\cos \omega t}{\alpha + j\omega}\, d\omega + \frac{j}{2\pi}\int_{-\infty}^{\infty} \frac{\sin \omega t}{\alpha + j\omega}\, d\omega \qquad [438]$$

The second of these two integrals remains proper in the limit $\alpha \rightarrow 0$, and hence one may set $\alpha = 0$ in the second term of Eq. 438 without further ado. This term is then the same as the second term in Eq. 434, and hence it is evident that the first term in Eq. 438 is supposed to yield the value $\frac{1}{2}$. The integral in this term, however, becomes improper in the limit $\alpha \rightarrow 0$ because the integrand then becomes infinite for $\omega = 0$. If, nevertheless, α is set equal to zero, one observes that the integrand is an odd function of ω, and since the limits of integration are symmetrical, the value of the integral, except for the difficulty in the vicinity of $\omega = 0$, should be zero. In other words, whatever value this integral may have must certainly be contributed by the immediate vicinity of the point $\omega = 0$. For this vicinity, which may be denoted by $-\rho < \omega < \rho$, one may again set $\cos \omega t$ equal to unity (since for very small increments from $\omega = 0$, $\cos \omega t$ differs from unity by a small quantity of the second order) and have for the first term of Eq. 438

$$\frac{1}{2\pi}\int_{-\rho}^{\rho} \frac{d\omega}{\alpha + j\omega} = \frac{1}{2\pi j}\ln\,(\alpha + j\omega)\Big]_{-\rho}^{\rho} = \frac{1}{2\pi j}\left[\ln\frac{\alpha + j\rho}{\alpha - j\rho}\right]_{\alpha \rightarrow 0} = \frac{1}{2} \qquad [439]$$

which is the correct value. However, the importance of retaining α up to the last step during this evaluation process should be noted. Until then, α plays an important role toward guiding the evaluation and preventing misinterpretations; after this step has been taken, α has served its purpose and may be retired without causing further difficulties.

The integration in Eq. 439 may be understood clearly if it is regarded as an integration in the complex plane. Thus, when

$$\zeta = \alpha + j\omega \qquad [440]$$

this integral takes the form

$$\frac{1}{2\pi j}\int_{(\alpha - j\rho)}^{(\alpha + j\rho)} \frac{d\zeta}{\zeta} \qquad [441]$$

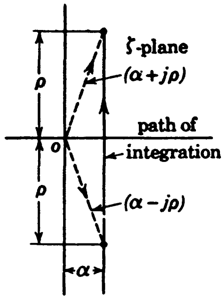


FIG. 33. The path of integration for the integral of Eq. 441 relevant to obtaining the unit step from its transform by the inverse Fourier integral.

The sketch in Fig. 33 indicates the path of integration in the complex ζ-plane. Inasmuch as the quantities $(\alpha + j\rho)$ and $(\alpha - j\rho)$ have the same magnitude, it is clear from the discussion of the logarithm function in Art. 17 of the preceding chapter that the value of the integral 441, the factor $1/2\pi j$ being omitted, is a pure imaginary quantity, equal in magnitude to the difference between the angle of $(\alpha + j\rho)$ and that of $(\alpha - j\rho)$. This net angle clearly approaches the value π as α is allowed to become zero, but it is to be noted that if α is set equal to zero to start with, the path of integration lies upon the imaginary axis in the ζ-plane, and the net angle might equally well be regarded as given by $-\pi$. With α retained, this ambiguity is avoided, and as soon as the correct interpretation has been seen, α may be discarded.

The second example to be considered is the function defined by

$$\begin{aligned} f(t) &= 0, \quad &&\text{for} \quad t < -\frac{\delta}{2} \\ f(t) &= 1, \quad &&\text{for} \quad -\frac{\delta}{2} < t < \frac{\delta}{2} \\ f(t) &= 0, \quad &&\text{for} \quad t > \frac{\delta}{2} \end{aligned} \qquad [442]$$

This is the same function which is considered in Art. 19 in the discussion of the approximation properties of the Fourier integral, as illustrated by the plot of Fig. 29. The transform, according to Eq. 351, is given by

$$g(\omega) = \frac{1}{2\pi}\int_{-\delta/2}^{\delta/2} e^{-j\omega t}\,dt = \frac{\delta}{2\pi}\left(\frac{\sin \omega \dfrac{\delta}{2}}{\omega \dfrac{\delta}{2}}\right) \qquad [443]$$

The similarity of this result to that given by Eq. 187 for the complex Fourier coefficients in the series representation of the function illustrated in Fig. 12 should be noted. Inasmuch as the latter represents the periodic repetition of the function considered at present, this similarity between the expression for the Fourier coefficients on the one hand, and the Fourier transform on the other is, of course, not surprising. It may be well to point out here that the symbol ω in Eq. 187 represents the *fundamental* angular frequency, and hence the quantity $n\omega$ in that equation (not just ω) becomes the analogue of ω in Eq. 443. The latter may be obtained

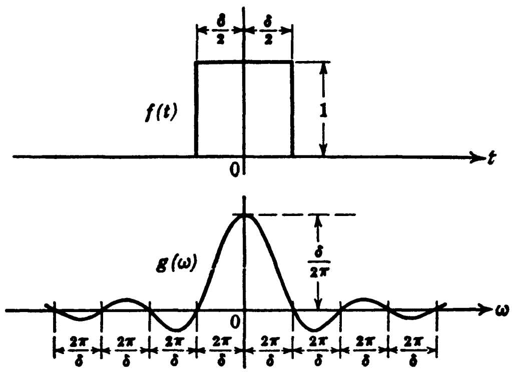


FIG. 34. The rectangular pulse of unit height and its associated Fourier transform.

from Eq. 187 through writing ω_1 for ω, and then applying the limiting process indicated in the expressions 349. With reference to Fig. 12, this process leaves the rectangular pulse at the origin but causes the adjacent pulses to move to infinity, thus yielding in the limit the function defined by the statement 442.

The transform $g(\omega)$ as a function of ω evidently has the appearance of the dotted curve of Fig. 13, showing the form of α_n as a function of $n\omega$. The important difference in the present example is the fact that the spectrum function $g(\omega)$ is continuous; that is, *all* frequencies are present, not just integer multiples of a fundamental frequency. Figure 34 shows both the time function $f(t)$ and the corresponding transform or spectrum function $g(\omega)$. This spectrum is a continuous one, whereas the spectrum shown in Fig. 13 is a line spectrum. It should be remembered, of course, that the amplitudes of $g(\omega)$ are not the harmonic amplitudes. The latter are all vanishingly small since their magnitudes are symbolically indicated by the differential notation $g(\omega)\,d\omega$. The function $g(\omega)$, nevertheless, shows how these amplitudes vary with ω for any increment $d\omega$ however small. This is true because, at any stage in the limiting process

indicated by the expressions 349, the differential spacing $d\omega$ of the " lines " in the continuous spectrum is constant, just as the finite spacing ω_1 (ω in Eq. 187) is in the corresponding periodic case.

In view of the statement 382 it is useful to observe that if the scale of ordinates in the plot of $g(\omega)$ in Fig. 34 is multiplied by 2π, the variables ω and t for these plots may be interchanged. That is, $2\pi g(t)$ may be regarded as the given time function and $f(\omega)$ as its transform. There is no need to reverse the algebraic sign of one of the interchanged variables ω or t in this example, because the functions are even. Thus, a time function like the curve for $g(\omega)$ in Fig. 34 is seen to have a spectrum like the curve for $f(t)$. This spectrum is, of course, also a continuous one, but the interesting feature about it is its finite extent.

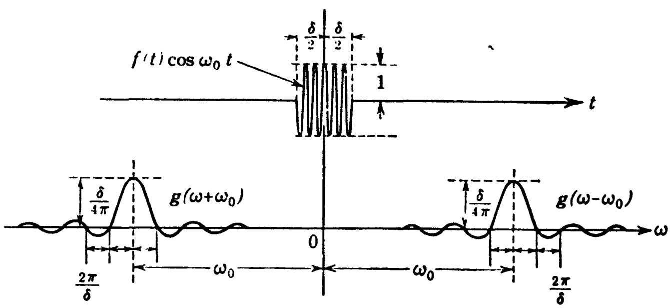


FIG. 35. A modulated cosine wave and its associated transform.

This result may be predicted on the basis of the discussion in Art. 18 regarding the resultant interference pattern of a frequency group. It is pointed out there that the frequency group of Fig. 27 has an envelope function like that plotted in Fig. 28 and expressed analytically by Eq. 339. As more and more lines are added to this frequency group, until it becomes a continuous spectrum of finite width like the function $f(t)$ in Fig. 34, the envelope function of the corresponding interference pattern approaches the form given by Eq. 341, which when plotted has the appearance of $g(\omega)$ in Fig. 34. In other words, the function $f(t)$ in Fig. 34, regarded as a spectrum function, is a continuous frequency group whose resultant interference pattern (time function) has the appearance of the function $g(\omega)$ in Fig. 34.

By means of the statement 385 it is a simple matter to determine the spectrum function which results when the time function of Fig. 34 is used to modulate the amplitude of a sinusoidal function of arbitrary frequency. Figure 35 illustrates such a time function, namely, a cosine function enclosed by a rectangular pulse. The same figure also shows the corre-

sponding transform, which consists of the linear superposition of the functions $g(\omega - \omega_0)$ and $g(\omega + \omega_0)$. These are obtained through multiplying the function $g(\omega)$ of Fig. 34 by ½ and displacing it respectively to the right and to the left of the origin by the amount ω_0, which is the angular frequency of the time function within the interval $-\delta/2 < t < \delta/2$.

It is interesting to consider the way in which the spectrum function for this example changes as the duration δ of the sinusoidal time function is increased. One should observe that the functions $g(\omega - \omega_0)$ and $g(\omega + \omega_0)$ consist essentially of a large hump whose amplitude is proportional to δ and whose width at the base of this large hump is inversely proportional to δ. As δ becomes very large, the spectrum function of Fig. 35 assumes the form of two very tall slim humps, one at ω_0 and the other at $-\omega_0$. In the limit $\delta \to \infty$ the spectrum function is given by two lines at the points $\pm\omega_0$, as one should expect from the fact that the single frequency ω_0 alone then characterizes the resulting time function.

A type of function very useful practically is readily derived from the function $f(t)$ in Fig. 34. If the amplitude of this function is set equal to $1/\delta$ instead of unity, the area enclosed by the rectangle equals unity no matter what δ may be. For a large value of δ, the rectangle is long and low, for a small value of δ, it becomes thin and tall. The transform $g(\omega)$ corresponding to this $f(t)$ is given by the expression 443 divided by δ, that is, by

$$g(\omega) = \frac{1}{2\pi}\left(\frac{\sin \omega \dfrac{\delta}{2}}{\omega \dfrac{\delta}{2}}\right) \qquad [444]$$

If δ is now allowed to approach zero, $f(t)$ degenerates into a single infinite ordinate at $t = 0$; that is, the rectangle has zero width and an infinite height but still encloses unit area. This limiting form of the function $f(t)$ of Fig. 34 is called a *unit impulse* and may be denoted by $u_0(t)$. Its transform is given by the expression 444 for the limit $\delta \to 0$. With this limiting value of the transform denoted by $v(\omega)$, it is readily seen that

$$v(\omega) = \frac{1}{2\pi} \qquad [445]$$

In other words, the transform of the unit impulse is a constant.

In view of the preceding discussion of the function $g_1(\omega)$ of Fig. 32 for the limiting process $\alpha \to 0$, this result may be arrived at in a different manner. The inverse transform of $g_1(\omega)$, namely $f_1(t)$ of Fig. 31, becomes equal to the constant ½ for $\alpha \to 0$; $g_1(\omega)$ degenerates into a single infinite ordinate at $\omega = 0$, although the area enclosed by the curve for

$g_1(\omega)$ remains constant and equal to $\frac{1}{2}$. According to the statement 382, the transform of $2\pi g_1(t)$ in the limit $\alpha \to 0$ is the constant $\frac{1}{2}$. Hence the transform of $2g_1(t)$ for $\alpha \to 0$, which is an impulse enclosing unit area, is the constant $1/2\pi$, the same as $v(\omega)$. In other words, the function $2g_1(t)$ for the limit $\alpha \to 0$ may be identified with the unit impulse $u_0(t)$, and its transform may be identified with $v(\omega)$ in Eq. 445. It is rather interesting that this conclusion should be true, in view of the fact that $u_0(t)$, in the argument of the preceding paragraph, is approached by the rectangular pulse $f(t)$ in Fig. 34, and the limiting form of $2g_1(t)$ is approached by the rounded pulse of Fig. 32.

The fact that the transform of the unit impulse is constant and equal to $1/2\pi$ may also be seen graphically. With reference to Fig. 34, if $f(t)$ is multiplied by $1/\delta$, then $g(\omega)$ is multiplied by $1/\delta$ also. Its amplitude is then independent of δ. As δ becomes smaller, the distance from the origin, $\omega = 0$, to the points $\omega = \pm 2\pi/\delta$ becomes larger. For a very small δ (tall narrow rectangular pulse) this distance is so large that, over a wide range of frequencies, $g(\omega)$ drops off very little from its value of $1/2\pi$ at $\omega = 0$. Finally, as δ approaches zero, the points $\omega = \pm 2\pi/\delta$ move to infinity, and the curve for $g(\omega)$ becomes a horizontal line $1/2\pi$ units above the ω-axis.

Conversely, one may let $g(\omega)$ in Fig. 34 approach the unit impulse. The area under this curve already equals unity because it represents the value of $f(0)$, according to the Fourier integral 350. Making use of the statement 382 again by saying that $(1/2\pi)f(\omega)$ is the transform of $g(t)$, and this time letting δ approach infinity, one finds that $g(t)$ approaches $u_0(t)$, and $(1/2\pi)f(\omega)$ approaches the constant $1/2\pi$.

The next function to be considered is shown in Fig. 36. This is the integral form $-\infty$ to t of the function $f(t)$ in Fig. 34 multiplied by $1/\delta$. Since the transform of the latter is given by Eq. 444, it follows from the statement 392 that the transform of the present time function is given by

$$g(\omega) = \frac{1}{2\pi j\omega}\left(\frac{\sin \omega \dfrac{\delta}{2}}{\omega \dfrac{\delta}{2}}\right) \qquad [446]$$

For small values of ω this transform behaves like the $g(\omega)$ of Eq. 424 for $\alpha = 0$, that is, like the transform of the time function defined by the relations 422 for $\alpha = 0$. In other words, the transform 446 for the function illustrated in Fig. 36 has a nonintegrable infinity at $\omega = 0$. This is due to the fact that the present time function does not fulfill the condition 352, and the same difficulty occurs as discussed in connection with the function defined by the relations 422 for $\alpha = 0$. Since the method of

dealing with this difficulty is now understood, however, the transform given by Eq. 446 may be accepted as an integrable function.

As δ is now allowed to approach zero, the function of Fig. 36 assumes the form shown in Fig. 37. This form, however, is the same as that of the function defined by the relations 422 and illustrated in Fig. 30 for the limit $\alpha \to 0$, as is also evident from the fact that the transform 446 for $\delta \to 0$ becomes identical with the transform 424 for $\alpha \to 0$. As indicated in Fig. 37, this limiting form of the function of Fig. 36 is denoted by the symbol $u_{-1}(t)$.

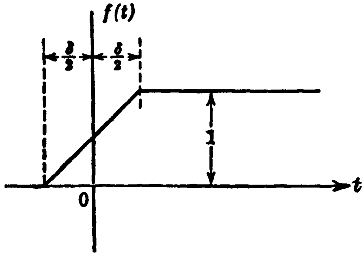


FIG. 36. The integral of a rectangular pulse similar to that of Fig. 34.

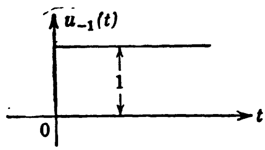


FIG. 37. The function of Fig. 36 where δ is allowed to approach zero.

Since the function of Fig. 36 is the integral from $-\infty$ to t of the $1/\delta$-multiplied time function of Fig. 34 for any value of δ however small, one may regard the function of Fig. 37 as representing the integral of the unit impulse $u_0(t)$. Symbolically this fact is expressed by

$$u_{-1}(t) = \int_{-\infty}^{t} u_0(t)\, dt \qquad [447]$$

in which the function $u_{-1}(t)$ is called the *unit step function* or, more briefly, the unit step. It is defined by the relations

$$\begin{aligned} u_{-1}(t) &= 0, \qquad \text{for} \quad t < 0 \\ u_{-1}(t) &= 1, \qquad \text{for} \quad t > 0 \end{aligned} \qquad [448]$$

and hence is identical with the function defined by the relations 422 for $\alpha = 0$. It plays an important part in the Heaviside Operational Calculus, but in the more recent expositions of this subject the unit impulse function $u_0(t)$ is found to be of greater value, chiefly because its transform is a constant.

From these discussions it becomes clear that the unit impulse may alternatively be expressed as the time derivative of the unit step, that is,

$$u_0(t) = \frac{du_{-1}}{dt} \qquad [449]$$

This is in agreement with the statement 389 since the transform for the unit step is given by Eq. 424 for the limit $\alpha \to 0$, and this result multiplied by $j\omega$ reduces to $1/2\pi$.

From a conservative mathematical point of view the differentiation of a function having a discontinuity is considered not permissible and is regarded as having no meaning. In view of the present discussion, however, it is clear that such operations are permissible provided they are properly interpreted. Thus it is possible to define functions and corresponding transforms for successive derivatives of the unit impulse. The first derivative is written

$$u_1(t) = \frac{du_0}{dt} \tag{450}$$

and its transform, according to the statement 389, is

$$j\omega v(\omega) = \frac{j\omega}{2\pi} \tag{451}$$

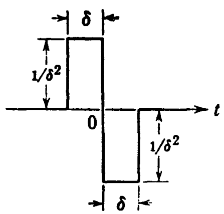


FIG. 38. A rectangular pulse doublet.

The function $u_1(t)$ is called the *unit doublet*. The reason for this name is clarified by reference to Fig. 38 which illustrates graphically the manner in which the derivation of this function is to be interpreted. Starting with the function shown in this figure, one obtains the unit doublet $u_1(t)$ by passing to the limit $\delta \to 0$. Since this limiting process may be indicated symbolically through replacing δ by the differential time increment dt, the correctness of this graphical interpretation may be seen analytically from the fact that

$$\frac{du_0}{dt} = \frac{u_0(t + dt) - u_0(t)}{dt} \tag{452}$$

or

$$\frac{du_0}{dt} = \frac{u_0\left(t + \frac{dt}{2}\right) - u_0\left(t - \frac{dt}{2}\right)}{dt} \tag{453}$$

When the statement 386 is applied, the corresponding transform is seen to be

$$\frac{v(\omega)e^{j\omega(dt/2)} - v(\omega)e^{-j\omega(dt/2)}}{dt} = \frac{v(\omega)j\omega\, dt}{dt} = j\omega v(\omega) \tag{454}$$

which agrees with the discussion surrounding Eqs. 450 and 451.

The unit doublet is evidently equivalent to a pair of equal but opposite impulses which are immediately adjacent to each other at the origin. The net effect may be likened to that of a *couple* in mechanics, and for this reason the term *couple* is sometimes used in place of the term *doublet* to designate the function $u_1(t)$.

It should be observed that the two impulses involved in this interpretation are not *unit* impulses. As shown in Fig. 38, the area enclosed by each rectangular pulse has the value $1/\delta$. In the limit $\delta \to 0$ this area becomes infinite, and the resulting impulse is seen to be one of infinite value rather than one of unit value. These considerations are pertinent to the proper interpretation of the expressions 452 and 453 inasmuch as u_0/dt symbolically represents an impulse of infinite value.

Through continuing in the same way a sequence of functions may be formed. The next in order is defined as

$$u_2(t) = \frac{du_1}{dt} = \frac{d^2u_0}{dt^2} \qquad [455]$$

and its transform, according to the statement 389, is given by

$$(j\omega)^2 v(\omega) = \frac{(j\omega)^2}{2\pi} \qquad [456]$$

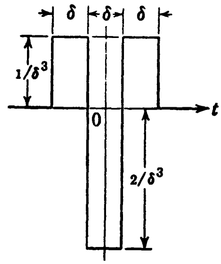


FIG. 39. A rectangular pulse triplet.

The function $u_2(t)$ may be interpreted graphically as the limit of the time function shown in Fig. 39 as δ is allowed to approach zero. It is equivalent to two equal but opposite doublets of infinite value centered about the origin and separated by the increment $\delta = dt$. Thus one may write

$$u_2(t) = \frac{du_1}{dt} = \frac{u_1\left(t + \frac{dt}{2}\right) - u_1\left(t - \frac{dt}{2}\right)}{dt} \qquad [457]$$

and, making use of the statement 386 again, have for the corresponding transform

$$\frac{j\omega v(\omega)e^{j\omega(dt/2)} - j\omega v(\omega)e^{-j\omega(dt/2)}}{dt} = (j\omega)^2 v(\omega) \qquad [458]$$

It is also possible to extend this sequence of functions in the opposite direction by successively integrating $u_0(t)$.

The first integration yields the unit function $u_1(t)$ or unit step. The next integration yields a time function which is linearly proportional to the time, like the current through a pure inductance when a constant

voltage is applied. In connection with practical problems there is little actual use for any of the functions in this sequence except the unit impulse and the unit step, although a recognition of the availability and the interpretation of the general sequence of functions together with their transforms proves to be a useful tool in the application of Fourier integral analysis to various practical problems.

According to the preceding discussion it should be clear that the transform of the general function $u_n(t)$ in this sequence is $(j\omega)^n/2\pi$. The sequence is referred to as the *singularity functions*, the unit impulse and the unit function being singularity functions of the order zero and minus one respectively. It should be observed that the singularity functions multiplied by 2π are the inverse transforms of the integer powers of $(j\omega)$. Inasmuch as the inverse transform of $(j\omega)^n$ cannot be found in the usual fashion because the integral 350 becomes improper, the discussion in this article may essentially be regarded as an interpretation process which avoids this difficulty and demonstrates the existence of such integrals under suitable limiting conditions.

25. The Error Function and the Sequence of Singularity Functions Based upon It

It has been seen from the preceding article that the unit impulse and its transform may be obtained through applying to other suitably chosen functions, besides the rectangular pulse, a limiting process by which they degenerate into a single infinite ordinate enclosing unit area. Correspondingly, the entire sequence of singularity functions and their transforms may be derived through applying suitable limiting processes to a variety of properly chosen functions. One of the most interesting of these is the so-called *error function* which has the form

$$f(t) = e^{-at^2} \qquad [459]$$

According to the integral 375, the transform of the error function is given by

$$g(\omega) = \frac{1}{\pi}\int_0^\infty e^{-at^2}\cos\omega t\,dt \qquad [460]$$

The integration yields

$$g(\omega) = \frac{e^{-\omega^2/4a}}{2\sqrt{\pi a}} \qquad [461]$$

Here it is, incidentally, interesting to observe that for the choice $a = \frac{1}{2}$, one has

$$f(t) = e^{-t^2/2} \qquad [462]$$

and

$$g(\omega) = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \tag{463}$$

In other words, the time function and its transform are identical except for a scale factor. In particular, if the transform is defined as $g^*(\omega)$, according to Eq. 365 and the Fourier integrals 366 and 367, the scale factor becomes unity and the time function is identical with its transform. This situation may also be achieved by using the integrals 369 and 370, for which the transform (considered as a function of the cyclic frequency $f = \omega/2\pi$) is defined in terms of $g(\omega)$ by Eq. 368. Thus, with a in Eqs. 459 and 461 equal to π, and Eq. 368 being used, the time function reads

$$h(t) = e^{-\pi t^2} \tag{464}$$

and its transform becomes

$$\hat{g}(f) = e^{-\pi f^2} \tag{465}$$

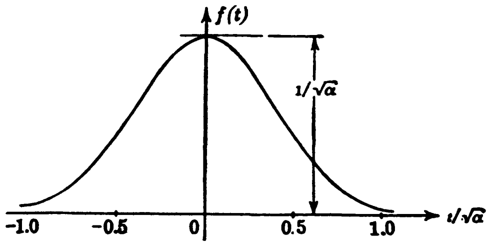


FIG. 40. The error function used in consideration of the singularity functions.

Returning now to the Eqs. 459 and 461, letting $a = \pi/\alpha$ and multiplying the resulting time function and its transform by the factor $1/\sqrt{\alpha}$, one has for the time function

$$f(t) = \frac{e^{-\pi t^2/\alpha}}{\sqrt{\alpha}} \tag{466}$$

and for its transform

$$g(\omega) = \frac{e^{-\alpha\omega^2/4\pi}}{2\pi} \tag{467}$$

This time function is plotted in Fig. 40. Since the area under the time function equals $2\pi g(0)$, it is clear that this area equals unity independent of the value of the parameter α. From Fig. 40 it is seen that as α becomes smaller the curve for $f(t)$ becomes taller and narrower and, in the limit $\alpha \to 0$, has the character of the unit impulse $u_0(t)$. At the same time one recognizes, from the form of Eq. 467, that $g(\omega)$ in the limit $\alpha \to 0$ becomes equal to the constant value $1/2\pi$, and hence identical with the transform

$v(\omega)$ of the unit impulse. Thus the first of the sequence of singularity functions is obtained from $f(t)$ of Eq. 466 for the limit $\alpha \to 0$, and its transform is found from Eq. 467.

The remainder of the singularity functions are obtained from the successive derivatives of the function $f(t)$ of Eq. 466 for the limit $\alpha \to 0$, and the corresponding transforms of this sequence of functions follow

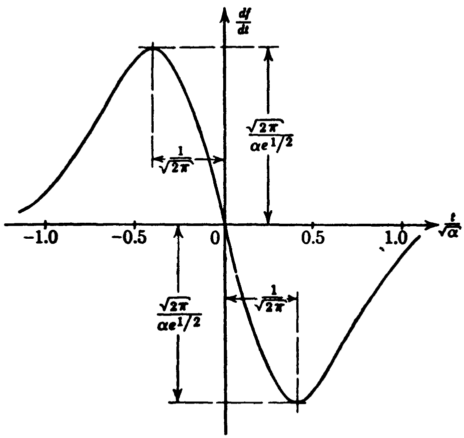


FIG. 41. The first derivative of the error function used in consideration of the unit doublet.

from the use of the statements 389 and 392 in connection with the transform 467. For example, the first and second derivatives of Eq. 466 read

$$\frac{df}{dt} = -\frac{2\pi t}{\alpha^{3/2}} e^{-\pi t^2/\alpha} \qquad [468]$$

and

$$\frac{d^2f}{dt^2} = \frac{2\pi(2\pi t^2 - \alpha)}{\alpha^{5/2}} e^{-\pi t^2/\alpha} \qquad [469]$$

whereas the corresponding transforms are given respectively by

$$g'(\omega) = j\omega g(\omega) = \frac{j\omega}{2\pi} e^{-\alpha\omega^2/4\pi} \qquad [470]$$

and

$$g''(\omega) = (j\omega)^2 g(\omega) = \frac{(j\omega)^2}{2\pi} e^{-\alpha\omega^2/4\pi} \qquad [471]$$

These time functions are shown in Figs. 41 and 42, from which it is clear that for $\alpha \to 0$ they have the general character of the singularity functions $u_1(t)$ and $u_2(t)$. In other words, the function 468 approaches

the unit doublet as α approaches zero, and the function 469 approaches the singularity function of order two.

It is interesting to compare the time functions of Figs. 34, 38, and 39, respectively, with those shown in Figs. 40, 41, and 42, and observe that, although the one set of curves is rectangular and the other set is smooth, both sets approach the same sequence of singularity functions when suitable limiting processes are carried out. This is evident from the fact that the transforms 467, 470, and 471 become identical with $v(\omega)$, $j\omega v(\omega)$, and $(j\omega)^2 v(\omega)$ in the limit $\alpha \to 0$.

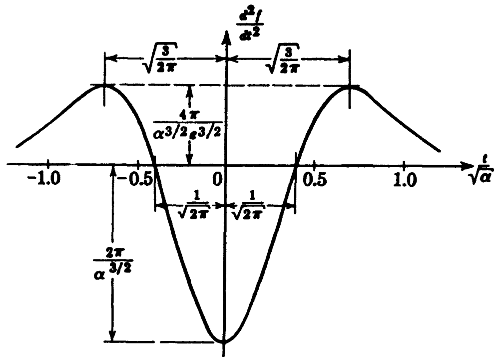


FIG. 42. The second derivative of the error function used in consideration of the singularity function of order two.

Since the error function and all its successive derivatives are smooth, use of them in the derivation of the singularity functions does not involve the mathematically doubtful steps encountered when the rectangular pulse and its derivatives are used for this purpose.

26. RELATION TO CONTOUR INTEGRALS

In applying the Fourier method of analysis to practical problems, one frequently encounters functions $g(\omega)$ which are quite complicated, and for which the evaluation of the synthesis integral 350 presents some difficulty. Although the function $g(\omega)$ may be complex, it should be clear that the integration is carried out with respect to the real variable ω, and hence is essentially an integration of a function of a real variable. Inspection of the integral representation of $g(\omega)$ according to Eq. 351, however, discloses that the variable ω occurs only in the exponent of the exponential function and is there associated with the operator j. This fact shows that the transform $g(\omega)$ may alternatively be regarded as a function of $j\omega$ and written in the form $g(j\omega)$. If this is done, and the

synthesis integral 350 is rewritten in the modified form,

$$f(t) = \frac{1}{j}\int_{-j\infty}^{j\infty} g(j\omega)\, d(j\omega)\, e^{j\omega t} \qquad [472]$$

it appears that this process of modification may be carried a step further through introducing the formal change of variable

$$\lambda = j\omega \qquad [473]$$

writing Eq. 472 in the form*

$$f(t) = \frac{1}{j}\int_{-j\infty}^{j\infty} g(\lambda)\, d\lambda\, e^{\lambda t} \qquad [474]$$

and, regarding $g(\lambda)$ as a function of a complex variable,

$$\lambda = \sigma + j\omega \qquad [475]$$

evaluating the integral 474 by the method of complex integration discussed in Art. 15 of the previous chapter. This method of integration, which is thus made available for the evaluation of the synthesis integral, holds promise for the simplification of many problems which present almost insurmountable difficulties unless one is exceptionally skilled in the art of real integration.

Before this method of dealing with the synthesis integral may be utilized, however, it is necessary to clarify several significant points which, in the above formal steps, are left in a somewhat doubtful state. First it is necessary to assure oneself that the complex transform $g(\lambda)$, which is obtained from the ordinary Fourier transform by the simple expedient of replacing $j\omega$ by a complex variable λ, is in fact the analytic continuation of the function $g(j\omega)$ into the complex domain. Such a justification is called for because the Fourier integral 351 defines the complex function g only in terms of the real variable ω, or one may say that it defines the function $g(\lambda)$ only for values of λ on the imaginary axis of the λ-plane. In other words, the Fourier integral 351 does not establish the existence of the function $g(\lambda)$ for all points in the λ-plane. In fact if one writes the integral 351 in the form

$$g(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, dt\, e^{-\lambda t} \qquad [476]$$

one immediately recognizes that this integral converges for $t < 0$ only for points in the left half of the λ-plane and for $t > 0$ only for points in the right half of this plane.

*Rewritten in terms of the variable λ, the Fourier integrals are commonly referred to as Laplace's integrals and $g(\lambda)$ as the Laplace transform of $f(t)$.

The existence of $g(\lambda)$ for all complex values of λ is, however, readily established with the help of the principle of analytic continuation and the uniqueness theorem for analytic functions (see Art. 11 of Ch. VI). Thus, suppose that a function $g(j\omega)$ is given (that is, the function $g(\lambda)$ for $\lambda = j\omega$) and let the problem be to find its analytic continuation into the λ-plane. Suppose furthermore that somehow a function $F(\lambda)$ of the complex variable λ is found which is identical with $g(\lambda)$ for all points on the imaginary axis (actually it is only necessary that $F(\lambda)$ and $g(\lambda)$ be identical for all points on an arbitrarily small portion of the imaginary axis or for an infinite number of discrete points having a limit point on the imaginary axis). Then, according to the uniqueness theorem, the functions $F(\lambda)$ and $g(\lambda)$ are identical everywhere in the λ-plane, and hence $F(\lambda)$ or $g(\lambda)$ is the desired analytic continuation of $g(j\omega)$.

Returning now to the integral 474, one finds that interpretation of it in terms of the method of complex integration raises a second pertinent question. The method of complex integration requires that the path of integration be in the form of a closed contour. According to the integral 474, the path is the entire imaginary axis, from minus infinity, through the origin, to plus infinity. Since infinity is regarded as a single point (this view is most easily appreciated through considering the complex plane replaced by its associated complex sphere), one may say that the closed contour requirement is met by the integral 474. A difficulty arises, however, because the integrand has an essential singularity at the point at infinity, since it contains the factor $e^{\lambda t}$. In the immediate vicinity of such a singularity a function is capable of assuming any assigned values (see Art. 13 of Ch. VI), and hence the method of passing through such a point cannot easily be disposed of. Unless the path of integration in the integral 474 is closed by passage through or around the point at infinity, the methods of complex integration cannot be applied, and yet the process of supplying this gap in the path of integration must be accomplished in a way which does not affect the value of the integral.

For the following argument it is necessary to assume that $g(\lambda)$ is a rational function and that, for large values of λ, it vanishes at least as $1/\lambda$. Modifications in the procedure which are called for when $g(\lambda)$ does not fulfill these conditions are more appropriately considered later. In the immediate vicinity of the point at infinity the integrand is then essentially represented by the factor $e^{\lambda t}/\lambda$. If the λ-plane is, for the moment, regarded as replaced by its associated complex sphere, according to the method of stereographic projection, one now contemplates by-passing the point at infinity by means of a path increment in the form of a small semicircular detour concentric with this point. In the ordinary λ-plane this detour corresponds to a semicircular path of very large radius with the origin as a center, as indicated in Fig. 43. This large

semicircle lies in the right or left half of the λ-plane according to whether the point at infinity is by-passed on the right or on the left.

In order that the detour shall contribute a negligible increment to the value of the integral, it is clear that $e^{\lambda t}/\lambda$ must at all events remain bounded for points on the detour. For such points it is also clear that λ has large values and that these values become infinite as the radius of the semicircular path about the point at infinity is made smaller and smaller, that is, as the radius R of the corresponding path in the λ-plane of Fig. 43 becomes larger and larger. In order for $e^{\lambda t}/\lambda$ to remain bounded, it is seen, therefore, that, for $t < 0$, the detour must lie in the right half plane where the real part of λ is positive and that, for $t > 0$, the detour must lie in the left half plane where the real part of λ is negative.

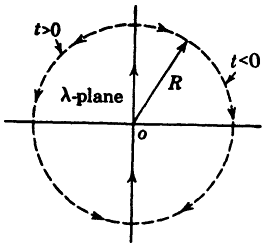


Fig. 43. The paths for replacing the Fourier integral along the imaginary axis by a contour integral.

It remains to show that, with this choice of detours for $t < 0$ and $t > 0$ respectively, the contributions to the integral 474 due to these added path increments are negligibly small. That is, one must show that the integral

$$I = \int_R \frac{e^{\lambda t}}{\lambda}\, d\lambda \tag{477}$$

extended over the semicircular paths as indicated in Fig. 43 has a negligible value for a sufficiently large value of R. For points on the semicircle

$$\lambda = Re^{j\theta} = R\,(\cos\theta + j\sin\theta) \tag{478}$$

and

$$\frac{d\lambda}{\lambda} = j\, d\theta \tag{479}$$

so that the integral 477 becomes

$$I = j\int e^{jRt\sin\theta}\cdot e^{Rt\cos\theta}\, d\theta \tag{480}$$

Since the factor $e^{jRt\sin\theta}$ has unit magnitude, it is clear that the value of the integral 480 extended over either of the two semicircular paths is certainly less than it would be if this factor in the integrand were omitted. As far as the magnitude of the integral is concerned, one may also drop the factor j and have

$$|I| < \int e^{Rt\cos\theta}\, d\theta \tag{481}$$

According to the choice of paths for $t < 0$ and $t > 0$, it is observed that for $t < 0$ the limits of integration are from $\theta = \pi/2$ to $\theta = -\pi/2$, and that for $t > 0$ they are from $\theta = \pi/2$ to $\theta = 3\pi/2$. For either path, the exponent $Rt \cos \theta$ is negative, and since $|\cos \theta| \leqq 1$, one may state that for $t < 0$,

$$|I| < \int_{\pi/2}^{-\pi/2} e^{Rt\cos\theta}\, d\theta < 2\int_0^{\pi/2} e^{-\frac{2}{\pi}R|t|\theta}\, d\theta = \frac{\pi(1 - e^{-R|t|})}{R|t|} \qquad [482]$$

and for $t > 0$,

$$|I| < \int_{\pi/2}^{3\pi/2} e^{Rt\cos\theta}\, d\theta < 2\int_0^{\pi/2} e^{-\frac{2}{\pi}R|t|\theta}\, d\theta = \frac{\pi(1 - e^{-R|t|})}{R|t|} \qquad [483]$$

For any nonzero $|t|$, the value of $|I|$ may, therefore, be made arbitrarily small through choosing a sufficiently large value of R.

According to the theory of contour integration, one recognizes that the radius R need be chosen only large enough so that the semicircular paths in Fig. 43 together enclose all the poles of the integrand. Since $e^{\lambda t}$ is an entire function, these poles are those of the rational function $g(\lambda)$.

As a simple illustrative example, let the time function $f(t)$ be the unit step $u_{-1}(t)$ with the transform $g(\lambda) = 1/2\pi\lambda$. The synthesis integral then reads

$$u_{-1}(t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{e^{\lambda t}}{\lambda}\, d\lambda \qquad [484]$$

In evaluating this contour integral one must again be reminded of the fact (discussed in Art. 24) that the transform of the unit step function is to be regarded as the limit of the function

$$g(\lambda) = \frac{1}{2\pi(\alpha + \lambda)} \qquad [485]$$

for $\alpha \rightarrow 0$. The pole of this function lies at the point $\lambda = -\alpha$, and since α is an arbitrarily small but nevertheless nonzero quantity, one observes that the pole of the integrand in the integral 484 must be regarded as lying, not at the origin of the λ-plane, but slightly to the left of this point. The closed contour for $t < 0$, therefore, does not enclose this pole, but the one for $t > 0$ does. Inasmuch as the residue of the integrand in this pole is unity (see Art. 15, Ch. VI, for the evaluation of residues), one readily recognizes that the integral 484 correctly represents the unit step function.

The necessity of writing the transform for the unit step in the form given by Eq. 485 and considering the limiting process indicated by $\alpha \rightarrow 0$ may be avoided if the path of integration in the vicinity of the origin is regarded as modified as shown in part (a) of Fig. 44. Instead of

passing through the origin, the path along the imaginary axis avoids the origin by passing to the right of it along a semicircular detour of vanishingly small radius. That a detour of this sort is necessary follows from inspection of the integral 484 inasmuch as a pole of the integrand lies upon the path of integration. However, unless one is aware of the limiting process which, in the limit, causes this pole to be located at the origin, one cannot know whether to by-pass this pole on the right or on the left. A knowledge of the limiting process which is necessary for the proper interpretation of the Fourier transform for the unit step function is thus seen to be necessary also for the removal of ambiguity in the reverse process of regaining this time function from its transform.

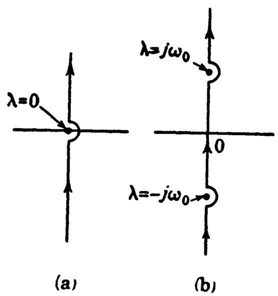


FIG. 44. Modification of the path of integration in the vicinity of poles of $g(\lambda)$.

It is possible for the integrand in the contour integral 474 to have several poles on the imaginary axis. A case of this kind arises when the time function has the form

$$f(t) = u_{-1}(t) \cdot \cos(\omega_0 t + \phi) \qquad [486]$$

which represents a steady sinusoid starting at $t = 0$. With the trigonometric function replaced by its exponential equivalent, the function 486 may be written

$$f(t) = u_{-1}(t) \cdot \tfrac{1}{2}\{e^{j\phi}e^{j\omega_0 t} + e^{-j\phi}e^{-j\omega_0 t}\} \qquad [487]$$

Utilizing the statement 385, one finds for the corresponding transform

$$g(\lambda) = \frac{e^{j\phi}}{4\pi} \cdot \frac{1}{(\lambda - \lambda_0)} + \frac{e^{-j\phi}}{4\pi} \; \frac{1}{\lambda + \lambda_0} \qquad [488]$$

in which

$$\lambda_0 = j\omega_0 \qquad [489]$$

The integrand in the contour integral for this $g(\lambda)$-function evidently has poles at the points $\lambda = \pm j\omega_0$ on the imaginary axis. In the evaluation of this contour integral, these poles must evidently be by-passed in the manner indicated in part (b) of Fig. 44. The residues of the integrand in these poles are seen to be respectively

$$\frac{e^{j\phi}}{4\pi} \cdot e^{\lambda_0 t} \quad \text{and} \quad \frac{e^{-j\phi}}{4\pi} \cdot e^{-\lambda_0 t} \qquad [490]$$

whence, observing again the selection of proper contours for $t < 0$ and

$t > 0$, one sees that the evaluation of the contour integral in this case correctly yields the time function 486 or 487.

In each of the two examples just discussed, one observes that $g(\lambda)$ does fulfill the condition of vanishing at least as strongly as $1/\lambda$ for large values of λ. It is appropriate at this point to consider in greater detail the necessity for this condition. One should here recall the discussion in Art. 8 relative to the character of Fourier coefficients for periodic functions which either are discontinuous themselves or possess discontinuities in their derivatives of the first or higher order. It is pointed out there that if the time function is discontinuous, the Fourier coefficients can become smaller no faster than $1/\nu\omega$, in which ν is the order of the harmonic coefficient and ω is the fundamental angular frequency. If the function is continuous but its first derivative is discontinuous, the coefficients can become smaller no faster than $1/\nu^2\omega^2$, and so forth.

Since these statements must obviously remain true as the period of the periodic function is made larger and larger, they apply also to the Fourier integral representation of a transient time function and its transform. The singularity functions $u_n(t)$, whose transforms are $1/2\pi\lambda^n$ are appropriate examples of this property. Inasmuch as the converse of these statements is evidently also true, one observes that the restriction that $g(\lambda)$ shall vanish for large λ at least as fast as $1/\lambda$ is seen to imply that the corresponding time function shall possess nothing worse than discontinuities. In other words, the method of contour integration is applicable to the evaluation of the synthesis integral only if the corresponding time function contains terms involving singularity functions of the order -1 or less. The impulse, for example, cannot be regained from its transform by the method of contour integration unless one employs special devices involving limiting processes similar to those used in the derivation of the transforms of such higher order singularity functions.

From a practical point of view this restriction on the function $g(\lambda)$ is hardly serious since the behavior of a physical system never exhibits the properties of an impulse or its derivatives unless the data are deliberately idealized. Moreover, one can, in such idealized cases, always apply a simple artifice to overcome the difficulty imposed by this restriction. For example, suppose $g(\lambda)$ approaches a constant value for large values of λ, implying that the corresponding time function contains an impulse. If the integrand is arbitrarily multiplied by $1/\lambda$, the statement 392 shows that the corresponding time function is replaced by its integral. Contour integration may then be applied, and the desired time function found through differentiating the result. In general, one may multiply the function $g(\lambda)$ by whatever power of $1/\lambda$ is needed to obtain the proper behavior for large values of λ and, after evaluating the contour integral, differentiate the result a corresponding number of times.

During this subsequent process of differentiation, one must observe the following precautions. If the time function resulting from the contour integration contains discontinuities, its derivative contains a corresponding number of impulses. For example, suppose a time function $f(t)$ has a discontinuity of the value h at $t = t_0$. The first derivative of $f(t)$ then contains the term $h \cdot u_0(t - t_0)$, its second derivative contains the term $h \cdot u_1(t - t_0)$, and so forth. Each discontinuity is treated in this manner whether it appears in the function $f(t)$ itself or in any of its subsequently formed derivatives. Besides terms of this sort, the derivatives of $f(t)$, of course, also contain terms representing the derivatives of the smooth portions of $f(t)$.

The second restriction which is placed upon $g(\lambda)$ in the above discussion of the resulting contour integration, namely, that $g(\lambda)$ be a rational function, may with some reservations next be relaxed to the extent of allowing $g(\lambda)$ to be a meromorphic function. As pointed out in Art. 18 of Ch. VI, this class of functions is more general than the rational ones in that the point at infinity may be an essential singularity. Thus $g(\lambda)$ is allowed to be a transcendental function, although its singularities in the finite λ-plane must still be ordinary poles.

This relaxation of the conditions imposed upon $g(\lambda)$ requires further consideration of the process of contour integration from two aspects. These are concerned, first, with the effect of the essential singularity at infinity and, second, with the possibility of an infinite number of poles in the rest of the λ-plane. A few simple examples will best illustrate how these matters may be dealt with.

Suppose the time function is the rectangular pulse defined by the relations 360. Its transform is given by

$$g(\lambda) = \frac{1}{2\pi}\int_{-\delta/2}^{\delta/2} e^{-\lambda t}\, dt = \frac{\sin \lambda \dfrac{\delta}{2}}{\pi\lambda} \qquad [491]$$

This is an entire transcendental function. Its only singularity is the one at infinity. Since the integrand in the contour integral 474 in this case has no poles at all, the entire process of evaluating this integral centers about the question of how the gap in the contour, which exists at the point at infinity, may be closed without affecting the value of this integral. Offhand, one may be tempted to conclude that the value of the integral is zero because the integrand has no poles. This conclusion is false, however, inasmuch as it is based upon the tacit assumption that the closure of the gap in the path of integration at infinity is to be dealt with in the manner described for rational $g(\lambda)$-functions. Such an assumption is presumptive.

The clue which leads one in the right direction is found through re-

writing the function 491 in the form

$$g(\lambda) = \frac{1}{2\pi\lambda}\left(e^{\lambda\delta/2} - e^{-\lambda\delta/2}\right) \tag{492}$$

whence the integrand in the integral 474 becomes

$$\frac{1}{2\pi\lambda}\left(e^{\lambda(t+\delta/2)} - e^{\lambda(t-\delta/2)}\right) \tag{493}$$

which, of course, is still an entire transcendental function. However, if the integrand is separated into two terms, each term is seen to have a simple pole at $\lambda = 0$, and in the vicinity of the point at infinity to behave in the same manner as already described for contour integrals involving rational $g(\lambda)$-functions except that the variable t is replaced respectively by $(t + \delta/2)$ and $(t - \delta/2)$. The paths which previously were chosen for $t < 0$ and $t > 0$ are now chosen for $(t \pm \delta/2) < 0$ and $(t \pm \delta/2) > 0$ respectively. Except for these changes, the integral for each term has the form of Eq. 484 for the unit step function. Hence one obtains the result

$$f(t) = u_{-1}\left(t + \frac{\delta}{2}\right) - u_{-1}\left(t - \frac{\delta}{2}\right) \tag{494}$$

which is recognized without difficulty to meet the definitions 360.

It should be observed that the question of how to close the gap in the path of integration at infinity is, in this example, resolved only after the integrand is separated into two terms, for the proper resolution with regard to one of these terms is different from that in the other. Unless the integrand is separated into two terms, it is obvious that the method of contour integration cannot be carried out for lack of an appropriate method of closing the path of integration. The method of contour integration can be applied to other $g(\lambda)$-functions having essential singularities at infinity only if similar artifices can be devised for dealing with this question.

Regarding the possibility of encountering transcendental $g(\lambda)$-functions having an infinite number of poles in the finite λ-plane, one may consider the example in which is sought the current response to a unit step voltage at the driving point of a lossless open-circuited transmission line. Except for a constant multiplier, the $g(\lambda)$-function in this case is of the form

$$g(\lambda) = \frac{\tanh\lambda}{\lambda} \tag{495}$$

which has simple poles at the points

$$\lambda_\nu = \pm\frac{j\nu\pi}{2} \qquad (\nu = 1, 3, 5, \cdots) \tag{496}$$

and an essential singularity at infinity.

The question of how to close the gap in the path of integration at infinity is disposed of through observing that $g(\lambda)$ remains bounded for large values of λ in the right or left half plane (which is not the case with the function 491 of the previous example). Hence large semicircles like the ones shown in Fig. 43 can be found on which the contributions to the contour integral for $t < 0$ and $t > 0$ are negligible. This question may, therefore, be resolved in a manner similar to that discussed for rational $g(\lambda)$-functions.

It remains to determine how one shall deal with the infinite number of poles of $g(\lambda)$ inasmuch as any semicircular path, no matter what its finite radius may be, cannot enclose all these poles. The difficulty presented by the fact that these poles lie upon the path of integration is, incidentally, overcome through avoiding them by means of small detours, after the fashion shown in Fig. 44. This procedure is valid since, for a transmission line with some loss, however small, the corresponding poles lie in the left half of the λ-plane.

The residues of the integrand in the integral 474 with the $g(\lambda)$-function of Eq. 495 are found to be given by

$$\rho_\nu = -\frac{\sinh^2 \lambda_\nu}{\lambda_\nu} e^{\lambda_\nu t} = \pm \frac{2}{\nu\pi j} e^{\pm j\nu\pi t/2} \qquad [497]$$

The significant point about this result is that the residues vary inversely as λ_ν. Hence they become smaller and smaller for poles which lie more and more remote from the origin of the λ-plane. The terms in the corresponding time function, therefore, become negligibly small for very remote poles. Since any number of terms in the desired time function are readily calculated, the question of choosing a sufficiently large radius R for the semicircular paths of Fig. 43 evidently depends upon the degree of approximation to which the result should be determined. The contour integral yields the time function in the form of an infinite series, which incidentally is recognized as being the Fourier series for a square wave.

By making use of the statement 392, and observing that multiplying $g(\lambda)$ by $1/\lambda^n$ has the effect of multiplying the corresponding residues by $1/\lambda_\nu^n$, one obtains a much more rapidly convergent series, but this effect is canceled by the subsequent n-fold differentiation. In practical problems in which the form of the resulting time function is not recognized from inspection of its series representation so easily as it is in the present example, a reasonable number of terms usually suffice for a sufficiently good approximation provided the series converges. The question of the convergence of the series can always be examined by means of the expression obtained for the residues.

In some practical problems one may also encounter $g(\lambda)$-functions

which are multivalued. For example, the determination of the response of an artificial transmission line to an applied unit step voltage leads to the transform

$$g(\lambda) = \frac{1}{2\pi\sqrt{\lambda^2 + \alpha^2}(\sqrt{\lambda^2 + \alpha^2} + \lambda)^n} \qquad [498]$$

in which n (the number of line sections) is an integer. The inverse transform is given by the integral

$$f(t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{e^{\lambda t}}{\sqrt{\lambda^2 + \alpha^2}(\sqrt{\lambda^2 + \alpha^2} + \lambda)^n}\, d\lambda \qquad [499]$$

The first step in the process of simplifying this integral is to let

$$\frac{\lambda}{\alpha} = w \qquad [500]$$

which converts 499 into the form

$$f(\alpha t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{e^{w\alpha t}}{\sqrt{w^2 + 1}(\sqrt{w^2 + 1} + w)^n}\, dw \qquad [501]$$

Next one introduces the change of variable indicated by

$$w = \frac{1}{2}\left(z - \frac{1}{z}\right) \qquad [502]$$

whence

$$\sqrt{w^2 + 1} = \frac{1}{2}\left(z + \frac{1}{z}\right) \qquad [503]$$

and, from the addition or subtraction of these two equations, one finds

$$z = \sqrt{w^2 + 1} + w \qquad [504]$$

and

$$\frac{1}{z} = \sqrt{w^2 + 1} - w \qquad [505]$$

Forming the differential of Eq. 504 yields

$$az = \frac{\sqrt{w^2 + 1} + w}{\sqrt{w^2 + 1}}\, dw \qquad [506]$$

whence, using Eq. 504 again, one has

$$\frac{dz}{z} = \frac{dw}{\sqrt{w^2 + 1}} \qquad [507]$$

By means of Eqs. 502, 504, and 507, the integral 501 assumes the greatly simplified form

$$f(\alpha t) = \frac{1}{2\pi j}\int_C \frac{e^{(z-1/z)\alpha t/2}}{z^{n+1}}\,dz \qquad [508]$$

in which the contour C in the z-plane must be chosen to correspond to the contour in the w-plane which is implied in the evaluation of the integral 501 by the method of complex integration.

Because of the multivalued character of the integrand in Eq. 501, the determination of an appropriate closed contour must be considered with some care. First it is observed that for large values of w the transform $g(w)$ varies as $1/w^{n+1}$, and hence, even for $n = 0$, it fulfills the requirements stated earlier regarding the method of closing the gap in the path of integration at the point at infinity. Hence one may again contemplate closing this gap in the manner shown in Fig. 43 for $t < 0$ and $t > 0$ without affecting the value of the integral.

Next it is seen that for integer values of n, the Riemann surface for the integrand of Eq. 501 has two leaves. The values $w = \pm j$ are branch points, and the portion of the imaginary axis between these is regarded as a branch cut. Since the point at infinity is not a branch point, the two leaves of the Riemann surface remain separate in this vicinity, and the path of integration around this point, therefore, remains on one of these leaves.

The branch points, which also are simple poles of the integrand (see the discussion immediately following Eq. 256 in Art. 18 of Ch. VI), must be avoided in the manner already described by means of small semicircular detours in the right half plane. In so doing, the path of integration also remains on the same leaf of the Riemann surface (as discussed in greater detail below) and the condition that the contour be closed is fulfilled.

In order to determine the contour C in the z-plane for the integral 508, it is necessary to consider in greater detail the substitution 502 and its inverse 504. For this purpose it is effective to determine the conformal map in the w-plane corresponding to orthogonal families of concentric circles and radial lines symmetrical with respect to the origin in the z-plane. With

$$z = re^{j\phi} \qquad [509]$$

the loci in the z-plane are defined by r = const and ϕ = const. Substituting into Eq. 502, one has

$$w = u + jv = \frac{r}{2}(\cos\phi + j\sin\phi) - \frac{1}{2r}(\cos\phi - j\sin\phi) \qquad [510]$$

whence

$$u = \frac{1}{2}\left(r - \frac{1}{r}\right)\cos\phi$$
$$v = \frac{1}{2}\left(r + \frac{1}{r}\right)\sin\phi \qquad [511]$$

Eliminating ϕ on the one hand and r on the other, one obtains respectively

$$\frac{v^2}{\frac{1}{4}\left(r + \frac{1}{r}\right)^2} + \frac{u^2}{\frac{1}{4}\left(r - \frac{1}{r}\right)^2} = 1 \qquad [512]$$

and

$$\frac{v^2}{\sin^2\phi} - \frac{u^2}{\cos^2\phi} = 1 \qquad [513]$$

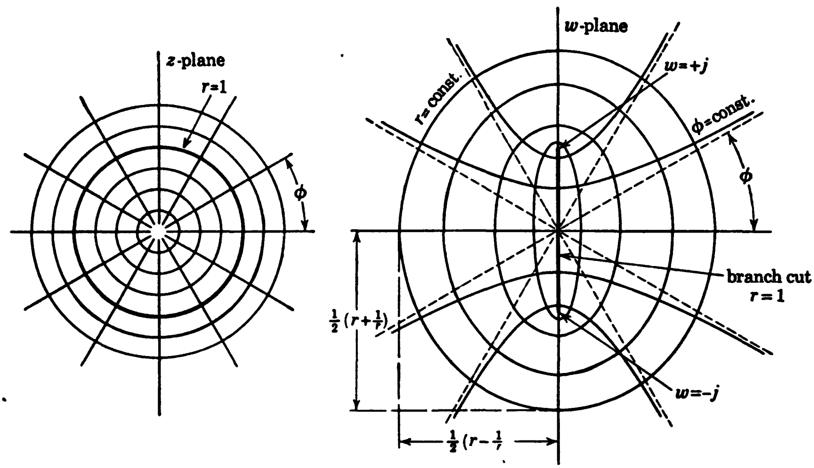


FIG. 45. Conformal representation of the substitution given by Eqs. 502 and 504.

For various values of r, Eq. 512 represents a family of confocal ellipses, with foci at the points $u = 0$, $v = \pm j$ (that is, $w = \pm j$). These loci are shown in Fig. 45. For $r = 1$, the ellipse degenerates into the doubly traversed portion of the imaginary axis between the points $w = \pm j$. The same ellipse is evidently obtained for reciprocal values of r, a very large or a very small value of r yielding a large ellipse with very little eccentricity, that is, one which is very nearly a circle concentric with the origin. One observes that the interior of the unit circle in the z-plane ($r < 1$) is mapped upon the entire w-plane and that the exterior of the unit circle in the z-plane ($r > 1$) is also mapped upon the entire w-plane.

The multivalued character of the function 504 is thus evident, inasmuch as two w-planes (the two leaves of a Riemann surface) are needed to map uniquely all points in the z-plane. The two leaves of the Riemann surface in the w-plane represent regions upon which the interior and the exterior of the unit circle in the z-plane are mapped respectively. The boundary between these two regions in the z-plane is the unit circle $r = 1$; in the w-plane it is the degenerate ellipse. The latter is the branch cut in the w-plane through which one passes from one leaf of the Riemann surface to the other.

For various values of ϕ, Eq. 513 represents a pair of families of confocal hyperbolas, which are images of each other about the real axis. These loci, which are also shown in Fig. 45, are orthogonal to the ellipses defined by Eq. 512, and their foci also lie at the points $w = \pm j$. The asymptotes of any hyperbola on one of the leaves of the Riemann surface make angles with the positive real axis which are equal to the values of ϕ appropriate to the corresponding branches of that hyperbola. On the other leaf, the angle between an asymptote and the positive real axis is $\pi - \phi$. Suppose the top leaf to be that one on which the angle is ϕ, and the lower one that on which the angle is $\pi - \phi$. Then if, for example, the branch of the hyperbola in the top leaf for $\phi = 30°$ (this is in the first quadrant) is traversed in the direction toward the branch cut, one finds, after following this hyperbola through the branch cut (into the second quadrant) that one is now on the lower leaf but still on an hyperbola to which $\phi = 30°$ is appropriate. If, instead of passing through the branch cut, one imagines jumping over it so as to remain on the top leaf, one finds oneself on an hyperbola (in the second quadrant) to which $\phi = 180° - 30° = 150°$ is appropriate.

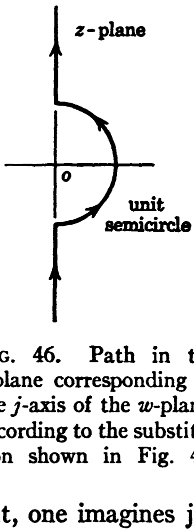


FIG. 46. Path in the z-plane corresponding to the j-axis of the w-plane, according to the substitution shown in Fig. 45.

With these properties of the substitution 502 in mind, it is easily appreciated that the path in the z-plane corresponding to the imaginary axis in the w-plane, traversed from $-j\infty$ to $+j\infty$ (avoiding the points $w = \pm j$ by remaining slightly to the right of them), has the form shown in Fig. 46. The process of closing the path in the w-plane by means of a large semicircle in the manner shown in Fig. 43 corresponds to closing the path of Fig. 46 in the z-plane in the same manner.* The large semi-

*It is true, of course, that a large circle concentric with the origin in the w-plane is not exactly (although very nearly) also a circle in the z-plane. This fact is, however, unimportant inasmuch as the argument regarding the closure of the path of integration does not require that the path increment in question be circular in form.

circles for $t < 0$ and $t > 0$ also lie in the right and left half planes respectively. The contour C in the integral 508 is thus determined.

For $t < 0$ the value of this integral is evidently zero, for the integrand has singularities only at the points $z = 0$ and $z = \infty$. For $t > 0$ the contour C encloses the one singularity at $z = 0$. This contour may, therefore, be given any other form, as long as it encloses the origin in the z-plane. Choosing the unit circle for the path C and writing in conformance with Eq. 509,

$$z = e^{j\phi} \tag{514}$$

one has

$$\frac{dz}{z} = j\,d\phi \tag{515}$$

and

$$\frac{1}{2}\left(z - \frac{1}{z}\right) = j \sin \phi \tag{516}$$

so that the integral 508 becomes

$$f(\alpha t) = \frac{1}{2\pi}\int_0^{2\pi} e^{j(\alpha t \sin \phi - n\phi)}\,d\phi \tag{517}$$

This result is seen to be identical in form with the integral representation for the Bessel function, as given by Eq. 307. Hence one has for $t > 0$,

$$f(\alpha t) = J_n(\alpha t) \tag{518}$$

It may be of interest to observe that the integral 508 alternatively represents the coefficients b_n in a Laurent expansion of the function $e^{(z-1/z)\alpha t/2}$ about its essential singularity at $z = 0$. One recognizes this fact from a comparison of Eq. 166 of Ch. VI with the integral 508, remembering that C is a contour enclosing the origin in the z-plane. As discussed in Art. 7, the substitution 514 converts the Laurent expansion into a complex Fourier series. One thus obtains the Fourier series representation for the function 303 dealt with in Art. 16.

An additional interesting feature about the transformation from the integral 501 to the integral 508 by means of the substitution 502 deserves special mention. In examining the integrand in the integral 501 one observes (as pointed out above) that the points $w = \pm j$ not only are branch points but also are simple poles, since the factor $\sqrt{w^2 + 1}$ becomes zero there. As the corresponding points in the z-plane, which from Eq. 504 are seen to be $z = \pm j$, the integrand in the equivalent integral 508, however, does not possess singularities. The reason for this peculiarity lies in the fact that the function $w(z)$ represented by the substitution 502 possesses saddle points (see Arts. 14 and 18, Ch. VI) at $z = \pm j$. This fact becomes evident upon the forming of the derivative of Eq. 502,

which reads

$$\frac{dw}{dz} = \frac{1}{2}\left(1 + \frac{1}{z^2}\right) \qquad [519]$$

This derivative has simple zeros for $z = \pm j$. Since the second derivative does not vanish at these points, they are saddle points of the first order. The vanishing of dw/dz at the points $z = \pm j$, corresponding to $w = \pm j$, may also be seen from Eq. 507. Since the quantity $dz/z = dw/\sqrt{w^2 + 1}$ remains regular in these points, the integral 508 does so likewise. For this reason it is unnecessary that the contour C of Fig. 46 be modified so as to avoid the points $z = \pm j$; the path of integration in the z-plane may pass through these saddle points.

It may also be of interest to recognize that the integral 508, except for the factor ½, is equivalent to Sommerfeld's integral 306. In place of the relation 514, one uses the substitution

$$z = e^{j\tau} \qquad [520]$$

in which τ is regarded as a new complex variable. Writing ρ for αt and dropping the factor ½, one has the integral

$$Z_n(\rho) = \frac{1}{\pi}\int_L e^{j(\rho \sin \tau - n\tau)}\, d\tau \qquad [521]$$

This is simply an alternative form for Sommerfeld's integral as given by Eq. 306. The latter form is obtained from Eq. 521 by making the additional change of variable

$$\zeta = \frac{\pi}{2} - \tau \qquad [522]$$

The minus sign resulting from the fact that $d\zeta = -d\tau$ is unimportant inasmuch as it may evidently be canceled through traversing in opposite direction the path of integration, which as yet is not specified.

For the various kinds of cylinder functions* defined by the integral 521, the path of integration L begins and ends at infinity. In order to insure the convergence of the integral, it is necessary (assuming $\rho > 0$) that the portions of L which extend toward infinity do so within regions of the τ-plane in which the real part of $j \sin \tau$ remains negative. Letting $\tau = \phi + j\eta$, one has

$$j \sin (\phi + j\eta) = -\cos \phi \sinh \eta + j \sin \phi \cosh \eta \qquad [523]$$

*For the demonstration showing that the function 521 formally satisfies Bessel's equation (which is dissociated from the present discussion), the reader is referred to the literature on this subject, for example, R. Courant and D. Hilbert, *Methoden der mathematischen Physik*, I (Julius Springer, 1924), 382, or E. T. Copson, *Theory of Functions of a Complex Variable* (Oxford, 1935), 313.

whence it is readily recognized that the regions in which the real part is negative are those shown cross-hatched in Fig. 47. Paths such as those labeled L_1 and L_2 may, therefore, be regarded as closing upon themselves at infinity, so that the principles of contour integration become applicable.

The detailed form of such a path within a cross-hatched region may, therefore, be modified at will without affecting the value of the integral. Thus the path L_1 may alternatively be assumed to lie along the imagi-

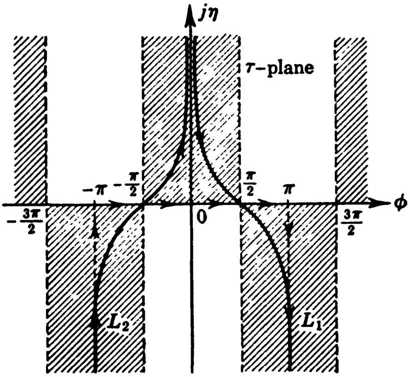


FIG. 47. Appropriate paths of integration passing through saddle points, used in the approximate evaluation of Sommerfeld's integral.

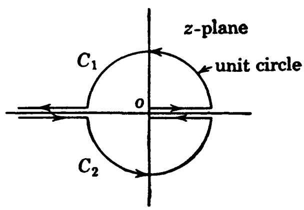


FIG. 48. Paths in the z-plane corresponding to modified versions of L_1 and L_2 in the τ-plane of Fig. 47.

nary axis from $\eta = \infty$ to $\eta = 0$, thence to lie along the real axis from $\phi = 0$ to $\phi = \pi$, and from there to proceed toward $\eta = -\infty$ along the vertical line $\phi = \pi$. Similarly, the path L_2 may be assumed to lie along the line $\phi = -\pi$ from $\eta = -\infty$ to $\eta = 0$, thence to lie along the real axis from $\phi = -\pi$ to $\phi = 0$, and finally to proceed toward $\eta = \infty$ along the imaginary axis.

The paths C_1 and C_2 in the z-plane, corresponding respectively to these modified versions of the paths L_1 and L_2 in the τ-plane, are shown in Fig. 48 (with due allowance for slight departures necessitated by drawing both paths in the same figure).

The specific functions defined by the integral 521 for the paths L_1 and L_2 are referred to as the *Hankel functions* or also as *Bessel functions* of the third kind. These are

$$H_n^{(1)}(\rho) = \frac{1}{\pi}\int_{L_1} e^{j(\rho \sin\tau - n\tau)}\, d\tau \qquad [524]$$

and

$$H_n^{(2)}(\rho) = \frac{1}{\pi}\int_{L_2} e^{j(\rho \sin \tau - n\tau)}\, d\tau \qquad [525]$$

For real as well as complex values of ρ, these functions are complex. The conjugate value of $H_n^{(1)}(\rho)$ for real values of ρ is given by

$$\bar{H}_n^{(1)}(\rho) = \frac{1}{\pi}\int_{\bar{L}_1} e^{-j(\rho \sin \tau - n\tau)}\, d\tau \qquad [526]$$

in which every point on the path $\bar{L}_1$ is the conjugate of a corresponding point on L_1, that is, $\bar{L}_1$ is the image of L_1 with respect to the real axis. If τ, the variable of integration, is replaced by $-\tau$, every point on the path of integration is replaced by its negative; that is, the path of integration becomes replaced by its image about the origin. If the latter path is denoted by $-\bar{L}_1$, one may write in place of Eq. 526

$$\bar{H}_n^{(1)}(\rho) = \frac{1}{\pi}\int_{-\bar{L}_1} e^{j(\rho \sin \tau - n\tau)}\, d(-\tau) \qquad [527]$$

It is now observed that the path $-\bar{L}_1$ is the image of L_1 about the imaginary axis. Reference to Fig. 47, therefore, shows that $-\bar{L}_1$ is identical with the path L_2 except for a reversal of the direction in which it is traversed. Changing this direction merely reverses the algebraic sign of the result. Hence one has

$$\bar{H}_n^{(1)}(\rho) = \frac{1}{\pi}\int_{L_2} e^{j(\rho \sin \tau - n\tau)}\, d\tau = H_n^{(2)}(\rho) \qquad [528]$$

that is, for real values of ρ, the Hankel functions of the first and second kind are conjugate complex.

The sum of these two functions may be expressed by a single integral of the form 521 in which the path L is the resultant of the paths L_1 and L_2. Reference to Fig. 48 shows that the corresponding resultant path in the z-plane is the unit circle enclosing the origin. According to the preceding discussion this path yields the Bessel function $J_n(\rho)$, hence

$$J_n(\rho) = \tfrac{1}{2}\{H_n^{(1)}(\rho) + H_n^{(2)}(\rho)\} \qquad [529]$$

For real values of ρ, one may regard the Bessel function as the real part of either of the Hankel functions. The relation 529, however, is by definition assumed to hold for complex as well as for real values of ρ.

For the sake of completing the present picture, it may be mentioned that the so-called cylinder functions of the second kind (also called Neumann functions) are given in terms of the Hankel functions by a relation complementary to Eq. 529, namely,

$$N_n(\rho) = \frac{1}{2j}\{H_n^{(1)}(\rho) - H_n^{(2)}(\rho)\} \qquad [530]$$

One observes that the Hankel functions are analogous to the exponential functions e^{jx} and e^{-jx}, the Bessel and Neumann functions to $\cos x$ and $\sin x$ respectively.

Asymptotic expressions valid for large values of the argument ρ may be obtained through evaluating the integrals 524 and 525 by the so-called "saddle-point" method. In terms of the variable τ defined by the substitution 520, the function 502 reads

$$w = j \sin \tau = u(\phi,\eta) + jv(\phi,\eta) \qquad [531]$$

The saddle points which, in the z-plane, occur for $z = \pm j$, are located in the τ-plane at the points $\tau = \pm\pi/2$. Reference to Fig. 47 shows that the paths L_1 and L_2 pass through these points. There the value of u is zero, whereas on either side of a saddle point u is negative. The exponential function

$$e^{j\rho \sin \tau} = e^{\rho(u+jv)} \qquad [532]$$

appearing in the integrals 524 and 525, has the magnitude

$$\left|e^{j\rho \sin \tau}\right| = e^{\rho u} \leqq 1 \qquad [533]$$

the value unity obtaining at a saddle point.

If one is mindful of the general character of the contours in the τ-plane defined by u = constant and v = constant, as discussed in Art. 14, Ch. VI (in particular, see Fig. 10 of Ch. VI for $s = 2$), one observes that if the path of integration is chosen to coincide with that contour v = constant which passes through the saddle point, the function u, and hence the exponential function given by Eq. 533, experience their most rapid rate of growth and subsequent decay. This fact is readily appreciated if one utilizes the analogy of regarding the loci u = constant as being contour lines in a mountainous terrain, and the orthogonal loci v = constant as indicating the direction of the gradient (direction of steepest ascent) in this terrain. The saddle point has the character of a mountain "pass," and the contour v = constant which passes through the saddle point represents the shortest route along which one may scale the height of the "pass" and descend into the valley beyond it. Clearly then, if this route is chosen as the path of integration, the values of the exponential function 533 pass continuously and most rapidly from the negligibly small magnitudes which obtain for points within the shaded regions of Fig. 47 remote from the origin, through their maximum (which occurs at the intersection of the path with the real axis), and back to negligibly small magnitudes again.

It is also readily appreciated that the portion of the path (in the vicinity of a saddle point) throughout which the exponential function has appreciable values becomes shorter as ρ becomes larger. For large values of ρ, therefore, the principal contribution to the value of either

of the integrals 524 or 525 is furnished by a rather short path increment in the immediate vicinity of the respective saddle point. An integration over such a short path increment alone then yields substantially the correct value of the desired function, the approximation becoming asymptotically better and better the larger the value of ρ.

In order to be able to carry out such an integration, one must first determine the direction (in the τ-plane) of the contour v = constant passing through the saddle point. From Eqs. 523 and 531 one has

$$u = -\cos\phi \sinh\eta$$

and [534]

$$v = \sin\phi \cosh\eta$$

At the saddle points $\tau = \pm\pi/2$ (that is, $\phi = \pm\pi/2$, $\eta = 0$), so that, as pointed out above, $u = 0$ and

$$v = v_0 = \Big[\sin\phi \cosh\eta\Big]_{\substack{\phi=\pi/2\\ \eta=0}} = \pm 1 \tag{535}$$

Along a contour v = constant one has

$$dv = 0 = \cos\phi \cosh\eta\, d\phi + \sin\phi \sinh\eta\, d\eta \tag{536}$$

from which

$$\frac{d\phi}{d\eta} = -\tan\phi \tanh\eta \tag{537}$$

By use of Eq. 535 the values of this derivative at the saddle points are found to be

$$\frac{d\phi}{d\eta} = -\sin\phi = \mp 1 \tag{538}$$

in which the minus sign applies to the saddle point at $\tau = +\pi/2$ and the plus sign to the one at $\tau = -\pi/2$. Hence the path L_1, which passes through the saddle point at $\tau = \pi/2$, should do so at an angle of -45 degrees with respect to the real axis, whereas the path L_2 should pass through its saddle point at $\tau = -\pi/2$ at an angle of $+45$ degrees. The paths shown in Fig. 47 comply with these conditions.

Next it is necessary to determine the detailed behavior of the function $u(\phi,\eta)$ in the immediate vicinity of the saddle points. This behavior must be expressed in terms of a variable which represents length measured along the respective paths of integration through these points. If this variable is denoted by s, the above determination of the path increments through the saddle points shows that one may write for the immediate

vicinities of these points

$$\tau = \pm\frac{\pi}{2} + se^{\mp j\pi/4} \qquad [539]$$

whence

$$w = j\sin\tau = \pm j\cos\left(se^{\mp j\pi/4}\right) \qquad [540]$$

Here s is regarded as a small quantity, so that one has approximately

$$w \cong \pm j\left(1 - \frac{s^2}{2}e^{\mp j\pi/2}\right) = \pm j\left(1 \pm j\frac{s^2}{2}\right) = -\frac{s^2}{2} \pm j \qquad [541]$$

and, therefore,

$$u = -\frac{s^2}{2} \qquad [542]$$

The value of v, according to the result 541, of course, agrees with that expressed by Eq. 535. The exponential factor 532 finally becomes

$$e^{j\rho\sin\tau} \cong e^{\pm j\rho}\cdot e^{-\rho s^2/2}, \qquad \text{for} \qquad |s| \ll 1 \qquad [543]$$

in which the plus sign in the exponent is to be used in the integral 524 and the minus sign in the integral 525.

The exponential factor $e^{-jn\tau}$, which also appears in these integrals, may, for the short path increment over which the integration is extended, be regarded as slowly variable compared with the factor 543, and hence as replaceable by its values at the saddle points. These are $e^{\mp jn\pi/2}$.

Observing, according to Eq. 539, that

$$d\tau = e^{\mp j\pi/4}\,ds \qquad [544]$$

one then has for the desired asymptotic expressions for the integrals 524 and 525

$$H_n^{(1)}(\rho) \cong \frac{1}{\pi}e^{j(\rho - n\pi/2 - \pi/4)}\int_{-\epsilon}^{+\epsilon} e^{-\rho s^2/2}\,ds \qquad [545]$$

and

$$H_n^{(2)}(\rho) \cong \frac{1}{\pi}e^{-j(\rho - n\pi/2 - \pi/4)}\int_{-\epsilon}^{+\epsilon} e^{-\rho s^2/2}\,ds \qquad [546]$$

in which 2ϵ is the length of the small path increment over which the integration extends. If the integrand $e^{-\rho s^2/2}$ is plotted versus s for large values of ρ, one finds that the area under this curve is confined substantially to the immediate vicinity of the origin ($s = 0$). In other words, the total area under this curve (which is obtained through integrating from $s = -\infty$ to $s = \infty$) differs from the area within a small region in

the vicinity of the origin by an amount which becomes smaller and smaller as ρ becomes larger and larger. The total area is given by

$$\int_{-\infty}^{\infty} e^{-\rho s^2/2}\, ds = \sqrt{\frac{2\pi}{\rho}} \tag{547}$$

Hence one has for the desired asymptotic expressions for the Hankel functions

$$H_n^{(1)}(\rho) \cong \sqrt{\frac{2}{\pi\rho}} \cdot e^{j[\rho-(2n+1)\pi/4]} \tag{548}$$

and

$$H_n^{(2)}(\rho) \cong \sqrt{\frac{2}{\pi\rho}} \cdot e^{-j[\rho-(2n+1)\pi/4]} \tag{549}$$

According to Eq. 529, a corresponding asymptotic form, valid for large ρ, is obtained for the Bessel function

$$J_n(\rho) \cong \sqrt{\frac{2}{\pi\rho}} \cos\left(\rho - \frac{2n+1}{4}\pi\right) \tag{550}$$

For $n = 0$ and $n = 1$ this result yields the formulas given by Eqs. 322 and 323 in Art. 16.

It may be well to point out that the approximate expressions just derived do not yield accurate results if the parameter n as well as the argument ρ is large, that is, if n and ρ are of the same order of magnitude. The truth of this statement is readily seen from the fact that if n is also large, the factor $e^{-jn\tau}$, appearing in the integrals 524 and 525, may no longer be regarded as essentially constant throughout the path increment over which the integration extends.

For a more general treatment of the present problem, one begins by setting

$$\rho = an \tag{551}$$

and rewriting the integrals 524 and 525 in the forms

$$H_n^{(1)}(an) = \frac{1}{\pi}\int_{L_1} e^{jn(a \sin\tau - \tau)}\, d\tau \tag{552}$$

and

$$H_n^{(2)}(an) = \frac{1}{\pi}\int_{L_2} e^{jn(a \sin\tau - \tau)}\, d\tau \tag{553}$$

In considering the saddle-point method of integration one then lets

$$w = j(a \sin\tau - \tau) = u + jv \tag{554}$$

The saddle points are those values of τ for which

$$\frac{dw}{d\tau} = j(a \cos \tau - 1) = 0 \tag{555}$$

yielding

$$\cos \tau = \frac{1}{a} \tag{556}$$

For large values of a (that is, for $\rho \gg n$), Eq. 556 yields very nearly

$$\cos \tau = 0 \qquad \text{or} \qquad \tau = \pm\pi \tag{557}$$

which are the saddle points of the function $j \sin \tau$ considered above.

The complete treatment of this problem (a being assumed real) requires separate consideration of the cases $a > 1$, $a = 1$, and $a < 1$. One obtains, in this manner, representations for the cylinder functions in the form of semiconvergent series of which the approximate results derived above are the first terms. The further detailed discussion becomes too specialized to be included under the heading of the present article,* in which the primary objective is to consider the essential principles involved in the use of complex integration for the evaluation of inverse Fourier transforms.

PROBLEMS

1. Make sketches of periodic functions which have the following specific characteristics:
   (a) The Fourier series contains only sine terms but all harmonics are present.
   (b) The Fourier series contains only sine terms and only odd harmonics.
   (c) The Fourier series contains only cosine terms but all harmonics are present.
   (d) The Fourier series contains only cosine terms and only odd harmonics.
   (e) The Fourier series contains sines and cosines but only odd harmonics.
   (f) The Fourier series has the property that the odd harmonics are sines and the even ones are cosines.
   (g) The Fourier series has the property that the odd harmonics are cosines and the even ones are sines.
   (h) The Fourier series has only even harmonics.
   (i) The Fourier series has harmonics of order α, 2α, 3α, $\cdots$, α being any fixed integer.

2. Given a set of functions $\phi_k(t)$ for $k = 1, 2, \cdots n$, with the following properties. $\phi_1(t)$ is periodic, having the period τ.

$$\phi_2(t) = \phi_1(t - t_1), \qquad \phi_3(t) = \phi_1(t - t_2), \qquad \cdots \phi_n(t) = \phi_1(t - t_{n-1})$$

in which t_1, t_2, $\cdots t_{n-1}$ are any finite quantities. Let the sum of these functions be denoted by $\phi(t)$, i.e.,

$$\phi(t) = \sum_{k=1}^{n} \phi_k(t)$$

*This material may be found, for example, in Courant-Hilbert, *op. cit.*, pp. 436–440, or E. T. Copson, *op. cit.* pp. 330–336.

If the function $\phi_1(t)$ has a Fourier series representation with the cosine and sine coefficients a_ν and b_ν, respectively, show that the corresponding coefficients for the resultant periodic function $\phi(t)$ are given by the expressions,

$$A_0 = na_0$$

$$A_\nu = a_\nu \sum_{k=1}^{n} \cos \nu\omega t_k - b_\nu \sum_{k=1}^{n} \sin \nu\omega t_k$$

$$B_\nu = b_\nu \sum_{k=1}^{n} \cos \nu\omega t_k + a_\nu \sum_{k=1}^{n} \sin \nu\omega t_k$$

Hint. Write the Fourier series for $\phi_k(t)$ and $\phi(t)$ in exponential form first and then convert to the trigonometric forms.

3. Suppose, for the set of functions defined in the preceding problem, one chooses $t_k = k(\tau/n)$. The periodic functions $\phi_1, \phi_2, \cdots \phi_n$ then form a cyclic group. Show that their sum $\phi(t)$, when it neither vanishes nor reduces to a constant, represents a periodic function with the period τ/n. Show that the same is true of the function $F(t) = \phi_1 \times \phi_2 \times \cdots \times \phi_n$ given by the product of the functions forming the cyclic group.

4. For the sum function $\phi(t)$ of the previous problem show that its Fourier coefficients A_μ and B_μ are given in terms of a_ν and b_ν for the component functions by the simple relationships: $A_\mu = na_\nu$, $B_\mu = nb_\nu$, for $\mu = 0, 1, 2, \cdots$ and $\nu = n\mu$. Correlate this result with the formulas given in Prob. 2.

5. As an application of the principles illustrated by the previous problems, let

$$\phi_1(t) = \sin \omega t \quad \text{for } 0 < t < \frac{\pi}{\omega}$$

$$\phi(t) = 0 \qquad \text{for } \frac{\pi}{\omega} < t < \frac{2\pi}{\omega}$$

and consider the cyclic groups for $n = 2, 3, 6, 12$, representing wave forms resulting from polyphase rectification. Compute the coefficients A_μ and B_μ according to the formulas of the previous problem and check, through carrying out the usual integration.

6. Square pulses of amplitude A and duration $\delta = \tau/10$ characterize the periodic function

$$\phi_1(t) = A \quad \text{for } 0 < t < \frac{\tau}{10}$$

$$\phi_1(t) = 0 \quad \text{for } \frac{\tau}{10} < t < \tau$$

Find the spectrum of this function and compare it with those obtained from a cyclic group for $n = 2$ and $n = 3$ according to the principles given in the previous problems.

7. In applying the formulas in Art. 15 to the numerical evaluation of Fourier coefficients for graphically given functions, a simplifying expedient is to alter the number of intervals n according to the order of the Fourier coefficient being calculated instead of using the same fixed number of intervals for the calculation of all coefficients. For example, if, in the use of formulas 299 to 302, one chooses $n = r$, all the α_{kr} and β_{kr} become ± 1 or zero, and the resulting computations are correspondingly simplified. This procedure, known as the Fischer-Hinnen method, must evidently be applied

with care. The results are not good for the fundamental and lower harmonics but improve with increasing order.

Try this method out on a function $f(x)$ whose values for $x = 0, \frac{\pi}{18}, \frac{\pi}{9}, \frac{\pi}{6}, \cdots \frac{\pi}{2}$ are respectively 0.0, 4.0, 7.0, 8.9, 9.7, 10, 9.5, 9.0, 8.7, 8.65, assuming that $f(-x) = -f(x)$ and $f(x + \pi) = -f(x)$. Compute harmonics through the seventh. Compare with results obtained from computations that do not utilize this simplified approach. Plot both results and compare with a plot of the given function.

8. For increments in x equal to $\pi/12$, starting with $x = 0$, the values of a given function $f(x)$ are:

0, 4.2, 8.9, 9.9, 9.5, 8.9, 8.5, 8.9, 9.5, 9.9, 8.9, 4.2, 0,
1.5, 1.8, 1.9, 2.1, 2.5, 3.0, 2.5, 2.1, 1.9, 1.8, 1.5, 0

Choosing $n = 12$, use the formulas 299 to 302 to compute the harmonics through the 11th. Plot the resulting partial sum and compare with the given function.

9. Given the function $f(x) = f(x + 2\pi k)$ defined by

$$f(x) = \begin{cases} \dfrac{x}{a} & \text{for } 0 < x < a \\[2ex] \dfrac{\pi - x}{\pi - a} & \text{for } a < x < 2\pi - a \\[2ex] -\dfrac{2\pi - x}{2\pi - a} & \text{for } 2\pi - a < x < 2\pi \end{cases}$$

Show that for this function $a_n = 0$ and

$$b_n = \frac{2 \sin na}{a(\pi - a)^2}$$

From this result determine the Fourier series for the following functions having the same period:

$$f_1(x) = \frac{\pi - x}{\pi} \qquad \text{for } 0 < x < 2\pi$$

$$f_2(x) = \begin{cases} \dfrac{2x}{\pi} & \text{for } 0 < x < \dfrac{\pi}{2} \\[2ex] \dfrac{2(\pi - x)}{\pi} & \text{for } \dfrac{\pi}{2} < x < \dfrac{3\pi}{2} \\[2ex] -\dfrac{2(2\pi - x)}{3\pi} & \text{for } \dfrac{3\pi}{2} < x < 2\pi \end{cases}$$

$$f(x) = \begin{cases} \dfrac{x}{\pi} & \text{for } 0 < x < \pi \\[2ex] -\dfrac{2\pi - x}{\pi} & \text{for } \pi < x < 2\pi \end{cases}$$

Make sketches of all the functions.

10. Find the Fourier expansion for a periodic function defined by

$$f(t) = \begin{cases} 0 & \text{for } 0 < t < a \quad \left(a < \dfrac{\tau}{2}\right) \\ \sin \pi \dfrac{t}{\tau} & \text{for } a < t < \tau - a \\ 0 & \text{for } \tau - a < t < \tau \end{cases}$$

Make a plot of the function and find the form of the Fourier series corresponding to $a = 0$. Find the sum of the series for $t = a$ and $t = \tau - a$.

11. A function $f(t)$ is defined by

$$f(t) = \begin{cases} 0 & \text{for } 0 < t < \dfrac{\tau}{2} - a \\ \cos \dfrac{\pi}{2a}\left(\dfrac{\tau}{2} - t\right) & \text{for } \dfrac{\tau}{2} - a < t < \dfrac{\tau}{2} + a \\ 0 & \text{for } \dfrac{\tau}{2} + a < t < \tau \end{cases}$$

Determine the Fourier series expansion. Plot the function for $a = \tau/5$ as well as the partial sums $s_1, s_2, \cdots s_5$, thus showing the manner in which the given function is approximated. Plot the spectrum function for the same value of a.

12. A periodic function consists of a regular succession of identical pulses of short duration (similar to the function of Prob. 11 for $a/\tau \ll 1$) the area under each pulse being A. Show that the values of the constant term and those of the fundamental and lower harmonic amplitudes are very nearly independent of the detailed pulse shape (whether rectangular, triangular, sinusoidal, etc.), being proportional only to the pulse area. Deduce the pertinent relationships. As the duration of the pulse is assumed to become shorter and shorter, the area remaining the same ($=A$), show that the Fourier coefficients are ultimately given by $a_0 = A/\tau$, $a_n = 2A/\tau$, independent of n.

13. Check the expansions of the following functions for the interval $0 \leqq x \leqq \pi$. Plot the functions and several terms of the series, noting rapidity of convergence.

$$\frac{\pi}{8} x(\pi - x) = \sum_{n=1,3,5\cdots}^{\infty} \frac{\sin nx}{n^3}$$

$$\frac{\pi}{96} (\pi - 2x)(\pi^2 + 2\pi x - 2x^2) = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\cos nx}{n^4}$$

14. Consider $f(x) = (\frac{1}{2})(\pi - x) \sin x$ within the interval $0 \leqq x \leqq \pi$, and check the following series representation:

$$\frac{1}{2} + \frac{1}{4} \cos x - \frac{1}{1\cdot 3} \cos 2x - \frac{1}{2\cdot 4} \cos 3x - \frac{1}{3\cdot 5} \cos 4x - \cdots$$

15. By using the Cauchy principle of convergence, show that the series

$$S = \sum_{n=1}^{\infty} \frac{\sin nx}{n}$$

converges uniformly except at the points $x = 0, 2\pi, 4\pi, \cdots$.

16. Determine the regions of uniform convergence of the series

$$S = \sum_{n=0}^{\infty} \frac{\cos nx}{n}$$

and define the points at which it diverges.

17. Show that the series

$$\sum_0^{\infty} z^n \cos nz \qquad \text{and} \qquad \sum z^n \sin nz$$

converge absolutely and uniformly inside a circle of unit radius.

18. Show that the series

$$\sum_0^{\infty} a_n z^n \cos nz \qquad \text{and} \qquad \sum a_n z^n \sin nz$$

converge absolutely and uniformly inside a circle of radius

$$R = \operatorname*{limit}_{n \to \infty} |a_n|^{-1/n}$$

19. Discuss the convergence of the series

$$S = \sum_{n=0}^{\infty} a_n x \sin nx$$

in which $a_n > a_{n+1} > a_{n+2} \cdots$ and $a_n \to 0$ for $n \to \infty$. Does the point $x = 0$ belong to the region of uniform convergence?

20. Show that the series

$$\sin x + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \cdots + \frac{\sin nx}{n} + \cdots$$

converges uniformly except at points $x = 0, \pm 2\pi, \pm 4\pi, \cdots$.

21. Show that the series

$$S = \sum_{n=1,3,5,\cdots}^{\infty} \frac{\cos nx}{n^2}$$

converges uniformly to the function $(\pi - 2x)\pi/8$ in the interval $0 \leqq x \leqq \pi$ and to the function $(x - 3\pi/2)\pi/4$ in the interval $\pi \leqq x \leqq 2\pi$.

22. Show that the expansion

$$\frac{1}{4}\left(\frac{\pi^2}{3} - x^2\right) = \cos x - \frac{1}{2^2}\cos 2x + \frac{1}{3^2}\cos 3x - \frac{1}{4^2}\cos 4x + - \cdots$$

converges uniformly in the interval $-\pi \leqq x \leqq \pi$. Through an obvious change of variable obtain the series

$$y^2 = \frac{c^2}{3} - \frac{4c^2}{\pi^2}\left[\cos\frac{\pi y}{c} - \frac{1}{2^2}\cos\frac{2\pi y}{c} + \frac{1}{3^2}\cos\frac{3\pi y}{c} - + \cdots\right]$$

and state its interval of uniform convergence.

23. Using the results of Prob. 22, show that

$$\frac{\pi^2}{12} = 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + - \cdots$$

and

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$

24. Using the results of Prob. 22, show that the series

$$x = \frac{c}{2} - \frac{4c}{\pi^2}\left[\cos\frac{\pi x}{c} + \frac{1}{3^2}\cos\frac{3\pi x}{c} + \frac{1}{5^2}\cos\frac{5\pi x}{c} + \cdots\right]$$

converges uniformly within the interval $0 \leqq x \leqq c$, and obtain the result

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots$$

25 Through the use of a Laurent expansion obtain the series

$$\frac{1 - r\cos z}{1 - 2r\cos z + r^2} = 1 + r\cos z + r^2\cos 2z + r^3\cos 3z + \cdots$$

in which r is real but z may be complex. Show that the region of convergence is defined as $-1 < r < 1$.

26. Find the continuous spectrum for the function

$$f(t) = \begin{cases} 0 & \text{for } t < 0 \\ Ae^{-\alpha t}\sin\omega_0 t & \text{for } 0 \leqq t < \infty \end{cases}$$

Plot the result for $A = 10$, $\omega_0 = 2\pi \times 10^6$, $\alpha = 2 \times 10^4$.

27. Find the Fourier transforms of the following functions:

(a) $$\begin{cases} f(t) = 0 & \text{for } t < 0 \\ f(t) = at & \text{for } 0 < t < t_1 \\ f(t) = at_1 - b(t - t_1) & \text{for } t_1 < t < t_2 = t_1\left(1 + \dfrac{a}{b}\right) \\ f(t) = 0 & \text{for } t_2 < t < \infty \end{cases}$$

(b) $$\begin{cases} f(t) = 0 & \text{for } t < t_1 > 0 \\ f(t) = e^{-\alpha(t-t_1)} & \text{for } t_1 < t < t_2 \\ f(t) = 0 & \text{for } t_2 < t < \infty \end{cases}$$

28. Let $G(\omega) = 2\pi g(\omega)$ and introduce $F(t) = e^{\sigma t} \times f(t)$ in which σ is a real quantity. Then letting $s = \sigma + j\omega$, obtain from the Fourier transforms the Laplace transforms

$$F(t) = \frac{1}{2\pi j}\int_{\sigma - j\infty}^{\sigma + j\infty} G(s)e^{st}\,ds$$

$$G(s) = \int_{-\infty}^{\infty} F(t)e^{-st}\,dt$$

Show that the condition for the existence of the transform $G(s)$ may be expressed by stating that the integral

$$\int_{-\infty}^{\infty} |e^{-\sigma t}F(t)|\,dt \quad \text{remains bounded}$$

Thus show that if $|F(t)| < Ce^{ct}$, in which C and c are positive real constants, the transform $G(s)$ exists so long as $\sigma > c$. (The quantity c is, therefore, called the abscissa of uniform convergence.)

29. Find the abscissa of uniform convergence for the following functions

$$\frac{1}{s},\ \frac{1}{s^2},\ \frac{1}{s^n},\ \frac{1}{s+a},\ \frac{1}{s-a},\ \frac{1}{(s+a)(s-b)},\ \frac{1}{(s+a)^2},\ \frac{1}{(s-a)^2},$$
$$\frac{s^3}{(s+a)(s-a)^2},\ \frac{1}{(\sqrt{s}+a)s},\ \frac{1}{\sqrt{s+a}\sqrt{s-b}},\ \frac{e^{st}}{s-a}$$

30. With the interpretation given in Prob. 28, and assuming $F(t) = 0$ for $t < 0$, show that the transforms of the following functions:

a, e^{at}, $\sin at$, $\cos at$, $\sinh at$, $\cosh at$, $e^{-at}\sinh bt$, $e^{-at}\cosh bt$, t, t^n, $t\sin at$, $t\cos at$ (in which a and b are positive real constants)

do exist; and by direct integration obtain for $G(s)$ respectively:

$$\frac{a}{s},\ \frac{1}{s-a},\ \frac{a}{s^2+a^2},\ \frac{s}{s^2+a^2},\ \frac{a}{s^2-a^2},\ \frac{s}{s^2-a^2},\ \frac{b}{(s+a)^2-b^2},$$
$$\frac{s+a}{(s+a)^2-b^2},\ \frac{1}{s^2},\ \frac{n!}{s^{n+1}},\ \frac{2as}{(s^2+a^2)^2},\ \frac{s^2-a^2}{(s^2+a^2)^2}$$

31. Through contour integration find $F(t)$ corresponding to

$$G(s) = \frac{1}{(s-a)(s-b)(s-c)}$$

in which a, b, c are real (unequal) quantities. Without further direct integration, what are the time functions corresponding to the following transforms:

$$\frac{s}{(s-a)(s-b)(s-c)},\ \frac{s^2}{(s-a)(s-b)(s-c)},\ \frac{1}{s(s-a)(s-b)(s-c)}$$

32. Starting with the pair of transforms

$$F(t) = e^{at} \qquad G(s) = \frac{1}{s-a}$$

find through convolution the time functions corresponding to

$$\frac{1}{(s-a)^2},\ \frac{1}{(s-a)(s-b)},\ \frac{1}{(s-a)(s+b)},\ \frac{1}{(s+a)(s+b)(s+c)},$$
$$\frac{e^{st}}{(s+a)(s-b)(s-c)}$$

Index

A

B

C